Internal Revenue Service (IRS) Record of 2013 Individual Tax Returns: How Big Data Can Shed Light on Tax Revenue

Masae Choi

Abstract— With advancements in all facets of technology, an explosion of data has followed. Enormous increases in the volume of data, the need for faster velocity, and ever increasing varieties of data formats have birthed the need for Big Data and its ecosystems. The key components of the Big Data ecosystems are tools, infrastructure, data sources, and processes such as data collection, storage, transformation, analysis and visualization of the useful information utilizing the Big Data. This project's goal is to use Cloudera's Hadoop Distributed File System (HDFS), Hive, and Pig to analyze and shed some light around tax revenue data. 2013 individual tax return data will be analyzed to answer the following questions:

Using the following adjusted gross income scales,

1 = $1 under $25,000
2 = $25,000 under $50,000
3 = $50,000 under $75,000
4 = $75,000 under $100,000
5 = $100,000 under $200,000
6 = $200,000 or more

How many tax returns filed for the above income scales?

What is the percentage of taxes that are collected for each of the above income scales?

What is the net effective rate of taxes paid for each of the above income scales?

What is the percentage of income that is generated, reflected in the tax return for each of the income scales?

To find answers to the questions above, Cloudera and the Hadoop Ecosystem provide a platform to query and analyze the data.

I. Introduction

Being that we are in an election year, many statistics are thrown around carelessly to aid various agendas. One argues that we are one of the highest tax paying countries in the world. Another argues that we do not pay enough in relation to other countries in the world. In both cases statistical data are used to prop up their arguments. Eventually, all comments just turn into rhetoric and leaves everyone uninformed and indifferent. Therefore, the hope is to utilize the Internal Revenue Service's 2013 individual tax return data set to get better analytic data to add just a little clarity in hopes that we all can move toward becoming more informed voters.

As we all know, anyone that generates revenue pays a portion of their income in taxes. The Internal Revenue Service (IRS) is the federal organization that is responsible for collecting those taxes. It also administers the Internal Revenue code which are laws governing activities around all types of taxes. According to Wikipedia, the first income tax was temporarily introduced to raise funds for the American Civil War at the rate of 3%. In 1913, the U. S.

Constitution changed to provision the permanence of income taxes as we know it today. Currently, the IRS collects around $2.4 trillion and processes 234 million tax returns per year.[1] Every election year, much is said around how much we should or should not pay in taxes. Generally, as far as taxes go, most want to pay as little as possible. However, most also agree that taxes should be collected to support basic services that are needed to run an effective system. So who is carrying the most burden in paying for these taxes? How much is being actually collected? The focus is to answer these questions and more using Big Data tools found in its ecosystem.

II. Overview of Big Data

In our current environment, enormous volumes of data are being collected, generated and stored due to advances in all forms of technology. From unstructured data to structured transactional data, data is being stored and analyzed in enormous volumes. Requirements to process these enormous volumes of data is bringing unique challenges to all facets of our organization. The concept of big data refers to these enormous volumes of data but also solution platforms that have been put together to solve the unique challenges that surround processing these large volumes of data. The following diagram shows the various tools and components around Big Data.

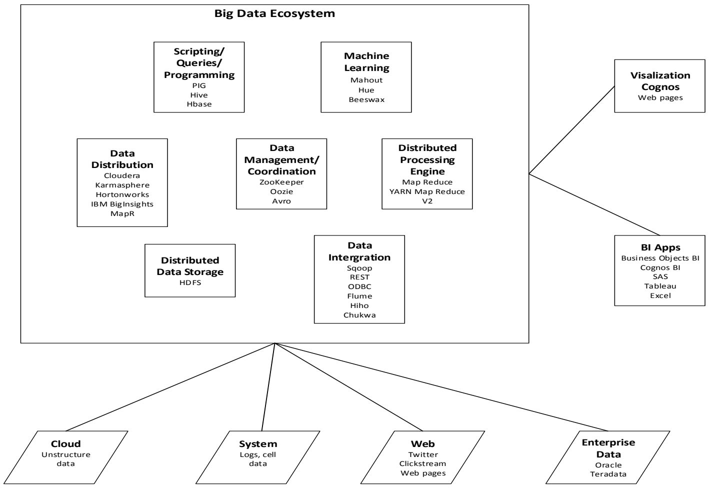


Figure 1. Big Data Ecosystem[2] referenced Big Data Components[3]

III. How are Big Data different than traditional technologies

Traditionally, databases that are largely utilized such as Oracle and MS SQL Server are relational based and centralized. They are designed to store structured data in rows and columns with a relational model in mind. Data relationships are at the foundation of this design, hence schema, fields and types are required for all data. The size of the data volume ranges from gigabytes to terabytes. For retrieval, SQL is utilized and performance can be an issue when performing detailed queries. On the other hand, Big Data is about handling the three Vs. They are volume, velocity and variety. In Big Data, volume ranges in petabytes to exabytes. To handle these enormous volumes, velocity, the speed in which big data are flowing through data processing and infrastructure, has to be optimized. Therefore, big data is fundamentally designed for distributed processing of the data. Lastly, Big Data can handle structured and unstructured data in a schema-less or flat schema.

IV. Overview of different Big Data Components

In Big Data, components can be categorized by the following segments.

Store – This component provides functionality to store vast amounts of data with an ability to handle a variety of data format one finds in the big data space. They are designed with a distributed data storage structure.

Management – Unlike relational designs where the tasks associated are in order, all these distributed tasks require their own management. This component provides that functionality.

Processing Engine - This component provides functionality to process large volumes of data in a distributed fashion to be able to process data efficiently.

Distribution: - Big Data is not just a single tool but rather a collection of components and associated tools. Distribution addresses the need to package this ecosystem for distribution.

Insight: This component provides a collection of machine language libraries to be able to obtain insight into the data. The intent is to utilize various algorithms or functions against big data for data analytics purposes.

Integration: This component provides functionality to communicate and transfer data with other systems.

Programming – This component provides functionality to code through a language to perform batch processing, tuning and performance around big data.

V. Hadoop Ecosystem Advantage

Following is a list of Hadoop ecosystem components and its description.[3]

- HDFS – Distributed Storage with high-throughput access
- MapReduce/YARN – Distributed processing of large volume of data on computer clusters
- HBase – Scalable, distributed database that supports structured data storage (NoSQL Database)
- Hive – data warehouse infrastructure that provides querying and summarization of data
- Mahout – Scalable Machine Learning and data mining library
- Pig – Scripting language for parallel processing
- ZooKeeper – Coordination for distributed services
- Ambari – Management & Monitoring
- Oozie – Workflow & Scheduling
- Sqoop/REST/ODBC – Data Integration

Hadoop's fundamental design seems to be based around how to facilitate processing and handling of vast amounts data in the most efficient way. This advantage becomes more apparent when processing big data. Hadoop's core design appears to be centered on efficient distributed processing of vast amounts of data. Hadoop slices the data and processes it in a parallel fashion and brings it at the end as a collection. This is apparent in HDFS which is a distributed data storage structure and MapReduce/YARN which is a distributed processing of data clusters.

VI. Data Definition

As a public organization, the Internal Revenue Service made their 2013 tax data available since The Statistics of Income (SOI) division bases its ZIP Code data on administrative records of individual income tax returns (Forms 1040) from the Internal Revenue Service (IRS) Individual Master File (IMF) system.[4] Included in these data are returns filed during the 12-month period, January 1, 2014 to December 31, 2014. The format that was used for this analysis was a CSV file with adjusted gross income classes. The figures in the data set is in thousands.

Each row of the data contains State, ZIP Code, AGI Class, total number of returns, various types of returns, exemptions, adjusted gross income, various types of income, various types of credit, deductions, and various taxes paid.

This 2013 tax return data set has been grouped by ZIP Code and AGI class. There are six classes and they are segmented using following definition:

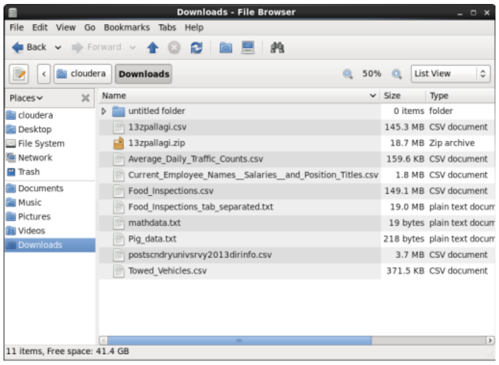
1 = $1 under $25,000
2 = $25,000 under $50,000
3 = $50,000 under $75,000
4 = $75,000 under $100,000
5 = $100,000 under $200,000
6 = $200,000 or more

This data had the following definitions and assumptions.[5]

- ZIP Code data are based on population data that was filed and processed by the IRS during the 2014 calendar year.
- State totals may not be comparable to State totals published elsewhere by SOI because of specific disclosure protection features in the ZIP Code data.
- Data do not represent the full U.S. population because many individuals are not required to file an individual income tax return.
- The address shown on the tax return may differ from the taxpayer's actual residence.
- State codes were based on the ZIP Code shown on the return.
- Excluded were tax returns filed without a ZIP Code and returns filed with a ZIP Code that did not match the State code shown on the return.
- Excluded were tax returns filed using Army Post Office (APO) and Fleet Post Office addresses, foreign addresses, and addresses in Puerto Rico, Guam, Virgin Islands, American Samoa, Marshall Islands, Northern Marianas, and Palau. ZIP Codes with less than 100 returns and those identified as a single building or nonresidential ZIP Code were categorized as "other" (99999).
- Income and tax items with less than 20 returns for a particular AGI class were combined with another AGI class within the same ZIP Code. Collapsed AGI classes are identified with a double asterisk (**).
- All numbers of returns variables have been rounded to the nearest 10.
- Excluded from the data are items with less than 20 returns within a ZIP Code.
- Excluded from the data are tax returns with a negative adjusted gross income.
- Excluded are tax returns representing a specified percentage of the total of any particular cell. For example, if one return represented 75 percent of the value of a given cell, the return was suppressed from the tabulation. The actual threshold percentage used cannot be released.

VII. Data Preparation

The public 2013 tax return dataset name 13zpallagi.csv was originally in .csv format. Therefore, direct upload to Cloudera was performed as shown below.



Name	Size	Type
untitled folder	0 items	folder
13zpallagi.csv	145.3 MB	CSV document
13zpallagi.zip	18.7 MB	Zip archive
Average_Daily_Traffic_Counts.csv	159.6 KB	CSV document
Current_Employee_Names__Salaries__and_Position_Titles.csv	1.8 MB	CSV document
Food_Inspections.csv	149.1 MB	CSV document
Food_Inspections_tab_separated.txt	19.0 MB	plain text docum
mathdata.txt	19 bytes	plain text docum
Pig_data.txt	218 bytes	plain text docum
postscndryunivsrvy2013dirinfo.csv	3.7 MB	CSV document
Towed_Vehicles.csv	371.5 KB	CSV document

11 items, Free space: 41.4 GB

Utilizing Metastore Manager, a table was created based on 13zpallagi.csv file.

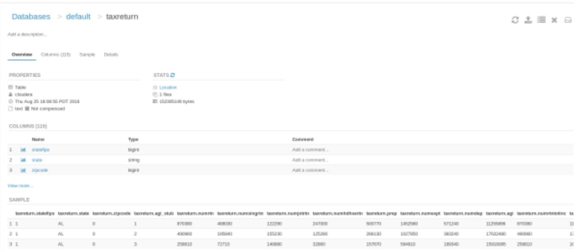


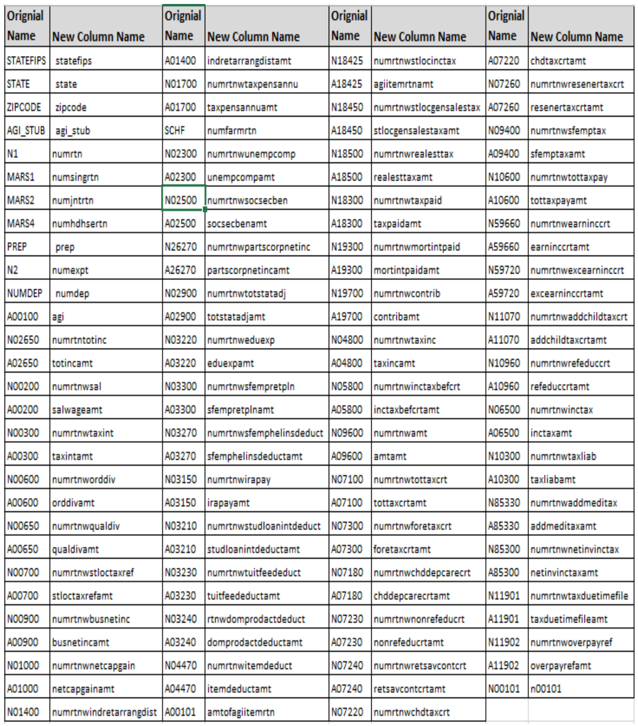
Table
cloudera
Thu Aug 25 18:08:55 PDT 2016
text Not compressed

1 files
152305149 bytes

	Name	Type	Comment
1	statefips	bigint	Add a comment...
2	state	string	Add a comment...
3	zipcode	bigint	Add a comment...

	taxreturn.statefips	taxreturn.state	taxreturn.zipcode	taxreturn.agi_stub	taxreturn.numrtn	taxreturn.numsingrtn	taxreturn.numjntrtn	taxreturn.numhdhsertn	taxreturn.prep	taxreturn.numexpt	taxreturn.numdep	taxreturn.agi	taxreturn.numrtntotinc
1	1	AL	0	1	870380	488030	122290	247000	500770	1452580	571240	11255898	870380
2	1	AL	0	2	490960	195840	155230	125200	266130	1027950	383240	17632480	490960
3	1	AL	0	3	258810	72710	146880	32860	157670	594810	180340	15916085	258810

The table resulted in 115 columns. Most of the column names were non-descriptive. (Ex. A26470, N03150) Therefore, based on the description of the columns, new column names were created for 115 columns. Following is the map of original column name to new.

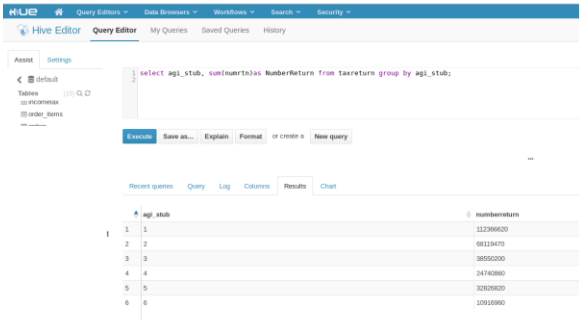


Orignial Name	New Column Name	Orignial Name	New Column Name	Orignial Name	New Column Name	Orignial Name	New Column Name
STATEFIPS	statefips	A01400	indretarrangdistamt	N18425	numrtnwstlocinctax	A07220	chdtaxcrtamt
STATE	state	N01700	numrtnwtaxpensannu	A18425	agiitemrtnamt	N07260	numrtnwresenertaxcrt
ZIPCODE	zipcode	A01700	taxpensannuamt	N18450	numrtnwstlocgensalestax	A07260	resenertaxcrtamt
AGI_STUB	agi_stub	SCHF	numfarmrtn	A18450	stlocgensalestaxamt	N09400	numrtnwsfemptax
N1	numrtn	N02300	numrtnwunempcomp	N18500	numrtnwrealesttax	A09400	sfemptaxamt
MARS1	numsingrtn	A02300	unempcompamt	A18500	realesttaxamt	N10600	numrtnwtottaxpay
MARS2	numjntrtn	N02500	numrtnwsocsecben	N18300	numrtnwtaxpaid	A10600	tottaxpayamt
MARS4	numhdhsertn	A02500	socsecbenamt	A18300	taxpaidamt	N59660	numrtnwearninccrt
PREP	prep	N26270	numrtnwpartscorpnetinc	N19300	numrtnwmortintpaid	A59660	earninccrtamt
N2	numexpt	A26270	partscorpnetincamt	A19300	mortintpaidamt	N59720	numrtnwexcearninccrt
NUMDEP	numdep	N02900	numrtnwtotstatadj	N19700	numrtnwcontrib	A59720	excearninccrtamt
A00100	agi	A02900	totstatadjamt	A19700	contribamt	N11070	numrtnwaddchildtaxcrt
N02650	numrtntotinc	N03220	numrtnweduexp	N04800	numrtnwtaxinc	A11070	addchildtaxcrtamt
A02650	totincamt	A03220	eduexpamt	A04800	taxincamt	N10960	numrtnwrefeduccrt
N00200	numrtnwsal	N03300	numrtnwsfempretpln	N05800	numrtnwinctaxbefcrt	A10960	refeduccrtamt
A00200	salwageamt	A03300	sfempretplnamt	A05800	inctaxbefcrtamt	N06500	numrtnwinctax
N00300	numrtnwtaxint	N03270	numrtnwsfemphelinsdeduct	N09600	numrtnwamt	A06500	inctaxamt
A00300	taxintamt	A03270	sfemphelinsdeductamt	A09600	amtamt	N10300	numrtnwtaxliab
N00600	numrtnworddiv	N03150	numrtnwirapay	N07100	numrtnwtottaxcrt	A10300	taxliabamt
A00600	orddivamt	A03150	irapayamt	A07100	tottaxcrtamt	N85330	numrtnwaddmeditax
N00650	numrtnwqualdiv	N03210	numrtnwstudloanintdeduct	N07300	numrtnwforetaxcrt	A85330	addmeditaxamt
A00650	qualdivamt	A03210	studloanintdeductamt	A07300	foretaxcrtamt	N85300	numrtnwnetinvinctax
N00700	numrtnwstloctaxref	N03230	numrtnwtuitfeededuct	N07180	numrtnwchddepcarecrt	A85300	netinvinctaxamt
A00700	stloctaxrefamt	A03230	tuitfeedeductamt	A07180	chddepcarecrtamt	N11901	numrtnwtaxduetimefile
N00900	numrtnwbusnetinc	N03240	rtnwdomprodactdeduct	N07230	numrtnwnonrefeducrt	A11901	taxduetimefileamt
A00900	busnetincamt	A03240	domprodactdeductamt	A07230	nonrefeducrtamt	N11902	numrtnwoverpayref
N01000	numrtnwnetcapgain	N04470	numrtnwitemdeduct	N07240	numrtnwretsavcontcrt	A11902	overpayrefamt
A01000	netcapgainamt	A04470	itemdeductamt	A07240	retsavcontcrtamt	N00101	n00101
N01400	numrtnwindretarrangdist	A00101	amtofagiitemrtn	N07220	numrtnwchdtaxcrt		

VIII. Data Analysis

To answer the above questions, the following queries were run against number of returns, total income tax amount, and total income amount.

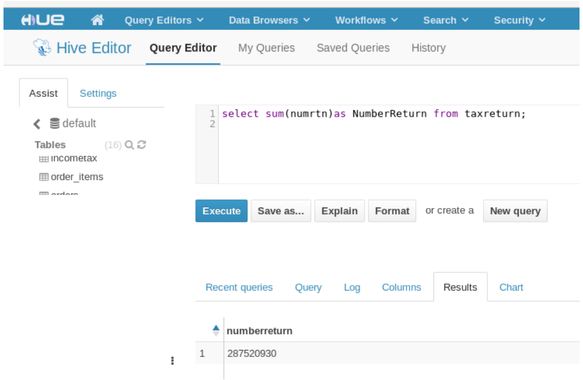
This query returned the number of returns filed by AGI Class.



```
select agi_stub, sum(numrtn)as NumberReturn from taxreturn group by agi_stub;
```

	agi_stub	numberreturn
1	1	112366620
2	2	68119470
3	3	38550200
4	4	24740860
5	5	32826820
6	6	10916960

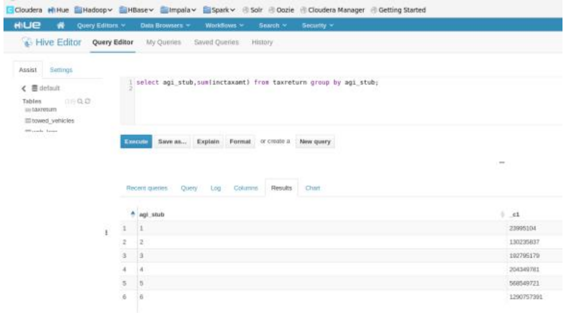
This query returned total number of returns for all the returns filed.



```
select sum(numrtn)as NumberReturn from taxreturn;
```

```
   numberreturn
1  287520930
```

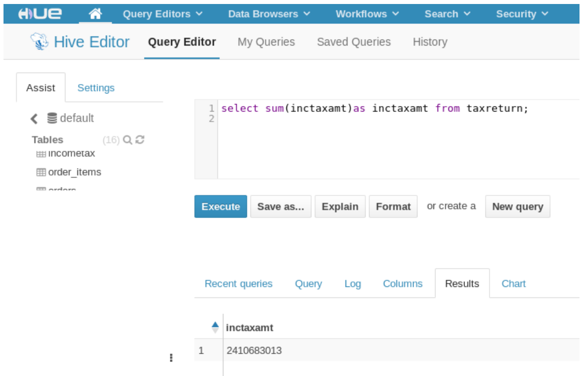
This query returned income tax amount by AGI Class.



```
select agi_stub,sum(inctaxamt) from taxreturn group by agi_stub;
```

```
   agi_stub   _c1
1  1          23995104
2  2          130235837
3  3          192795179
4  4          204349781
5  5          568549721
6  6          1290757391
```

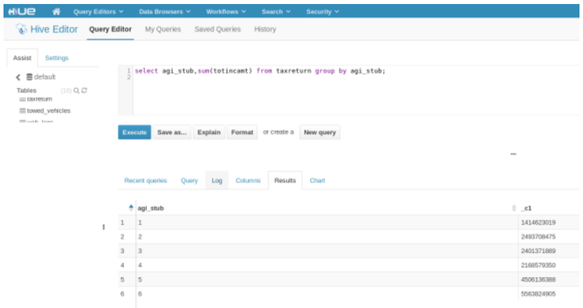
This query returned total income tax amount for all the returns filed.



```
select sum(inctaxamt)as inctaxamt from taxreturn;
```

```
   inctaxamt
1  2410683013
```

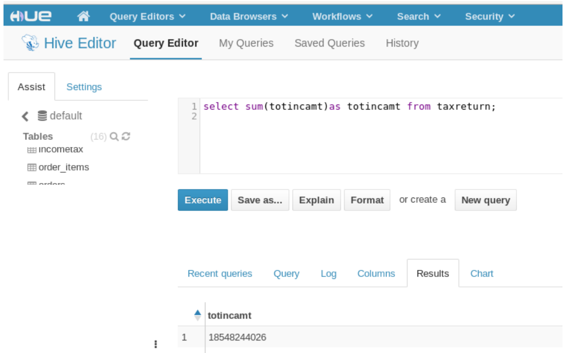
This query returned total income amount by AGI Class.



```
select agi_stub,sum(totincamt) from taxreturn group by agi_stub;
```

```
   agi_stub   _c1
1  1          1414623019
2  2          2493708475
3  3          2401371889
4  4          2168579350
5  5          4506136388
6  6          5563824905
```

This query returned total income amount for all the returns filed.



```
select sum(totincamt)as totincamt from taxreturn;
```

```
   totincamt
1  18548244026
```

After obtaining all the query results, the percentage of the AGI against the total calculation was performed for all three total queries by AGI. Secondly, by dividing income tax amount by total income amount for each AGI class, net tax return was calculated.

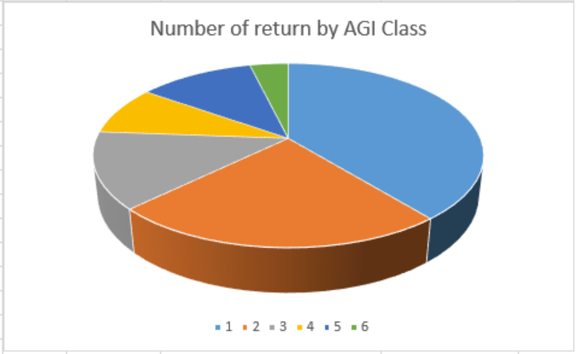
The following are the results of the calculation:

These are combined total query results and percentage calculation results.

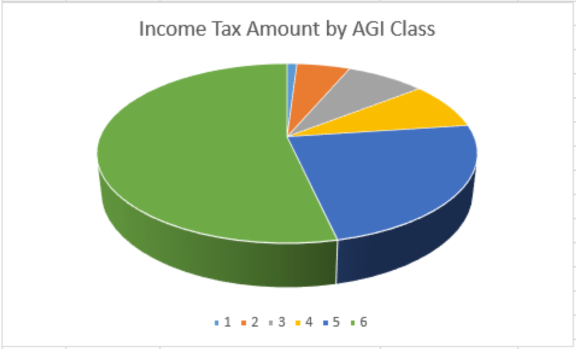
AGI Class	Sum of NumReturn	Sum of Income tax amount	Sum of Total income amount
1 = $1 under $25,000	112,366,620,000	$ 23,995,104,000	$ 1,414,623,019,000
2 = $25,000 under $50,000	68,119,470,000	$ 130,235,837,000	$ 2,493,708,475,000
3 = $50,000 under $75,000	38,550,200,000	$ 192,795,179,000	$ 2,401,371,889,000
4 = $75,000 under $100,000	24,740,860,000	$ 204,349,781,000	$ 2,168,579,350,000
5 = $100,000 under $200,000	32,826,820,000	$ 568,549,721,000	$ 4,506,136,388,000
6 = $200,000 or more	10,916,960,000	$ 1,290,757,391,000	$ 5,563,824,905,000
Grand Total	287,520,930,000	$ 2,410,683,013,000	$ 18,548,244,026,000

AGI Class	% of Return Filed	% of IncomeTax Amt	% of Total Income Amt	Net tax return %
1 = $1 under $25,000	39%	1%	8%	2%
2 = $25,000 under $50,000	24%	5%	13%	5%
3 = $50,000 under $75,000	13%	8%	13%	8%
4 = $75,000 under $100,000	9%	8%	12%	9%
5 = $100,000 under $200,000	11%	24%	24%	13%
6 = $200,000 or more	4%	54%	30%	23%

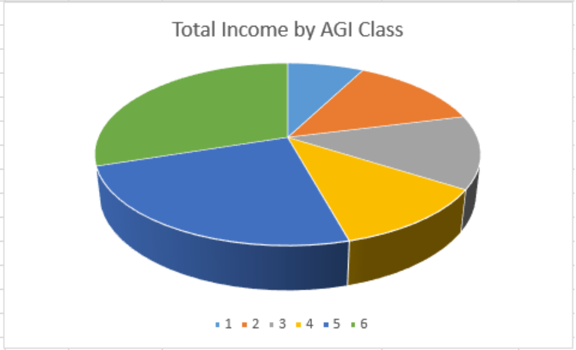
The graph shows number returns by AGI Class.



The graph shows income tax amount by AGI Class.



The graph shows total Income by AGI Class.



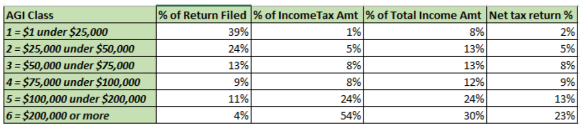
IX. Results/Observations

After performing the analysis, the following chart was created. This chart answered all of the following questions:

Questions:

- *How many tax returns filed for the above income scale?*
- *What is the percentage of taxes that are collected for each of the above income scales?*
- *What is the net effective rate of taxes paid for each of the above income scales?*
- *What is the percentage of income that are generated reflected in the tax return for each of the income scales?*

Answer:



AGI Class	% of Return Filed	% of IncomeTax Amt	% of Total Income Amt	Net tax return %
1 = $1 under $25,000	39%	1%	8%	2%
2 = $25,000 under $50,000	24%	5%	13%	5%
3 = $50,000 under $75,000	13%	8%	13%	8%
4 = $75,000 under $100,000	9%	8%	12%	9%
5 = $100,000 under $200,000	11%	24%	24%	13%
6 = $200,000 or more	4%	54%	30%	23%

Furthermore, the analysis revealed quite a number of interesting observations. The first interesting observation noted was the fact that 76% of the population that filed returns in 2013 made AGI less than $75,000. On the flip side, 15% of the population that filed returns in 2013 made more than $100,000 of AGI. The second shocking observation was that the AGI groups, (1, 2, and 3), which consists of 75% of the population paid 14% of the total income tax collected for 2013. On the other side of the spectrum, the top two AGI group, (5 and 6), which consists of 15% of the population paid 78% of the total income taxes collected in 2013. The fact that 76% of the population paid 15% of the total taxes and that the top 14% paid 78% of the total taxes for 2013 shows the 80-20 principle at work. The last observation was that the top two AGI group, (5 and 6), which consists of 15% of the population, made 54% of the total income identified through income tax filed in 2013. This shows that the top two AGI makes more income than rest of AGI groups combined for 2013.

X. Conclusion

Based on the analysis interesting observations were found. The fact that 76% of the population paid 15% of the total taxes and that the top 14% paid 78% of the total taxes for 2013 showed which segment of the population was carrying the most burden for total tax collected. However, it also showed that 76% made less than $75,000 making the tax burden hard to carry for most. When looking at the chart created, one can see that AGI group one with less than 25K AGI pays 2% of taxes. Therefore, one political candidate can use these statistic to make a statement that 39% of the population pays only 2% taxes. However, in my opinion that would be distorting the view of the situation. Since removing important fact that the group only made $25K, does change the implication of the statement.

Personally, by performing these analyses against 2013 data, I was able to get more clarity around taxes that are collected. Therefore, next time statistics are thrown around carelessly to aid various agendas, I will be more equipped to decipher facts from lies and can move toward becoming a more informed voter.

XI. References

[1] Wikipedia, the free encyclopedia (2016). Internal Revenue Service [online] Available: en.wikipedia.org/wiki/Internal_Revenue_Service

[2] Choi, Masae (2016) CIS 436 - Assignment 1, p4.

[3] Krishnan, Krish (2013) Data Warehousing in the age of big data (pp. 53-54), Waltham, Massachusetts: Morgan Kaufmann

[4] Data.Gov (2016). Zip Code data: tax year 2013 [online] Available: data.gov

[5] Data.Gov (2016). Zip Code data: tax year 2013, *Tax year 2013 Documentation Guide Available*, [online] data.gov

[6] Kakade, Sunilkumar (2016) CIS Big data management and analytics: Session 1 (p35)