Introduction: The data provided is a typical bookkeeping data set held by a population update based on Region by considering the sample sex, Race and getting the insights of the income, Income of the family members according to size.

Here In document, I consider few categorical variable (HAVING_HEALTHPLAN, MARSTAT_KEY, SAMPLE_SEX, COUNTRY_OF_BIRTH) and numerical variable (NET_WORTH_, INCOME_, YEAR_OF_BIRTH). These numerical and categorical data used describe the complete parameter and statistics of the given data.

After loading the dataset, we perform the cleaning step. It comprises of primarily removing data records which have irrelevant data (this could include incompatible data types, outliers, extreme values). For these values if found in dataset we replace them with the mean of the dataset.

The first section of analysis is descriptive statistics. Descriptive tables for dataset considering categorical variables, numerical variable and exploratory data analysis are shown followed by inferential analysis using histogram, boxplots and scatterplots.

ANALYSIS:

1.DESCRIPTIVE STATISTICS

1.1 Statistic descriptive table:

Table 1: Summary stats of Dataset - Analyzed data using below data fields.

	DATASET				
	N	**Mean**	**Max**	**Min**	**SD**
ID	10251	5862.310409	12679	2	3463.713809
YEAR	10251	1990	1990	1990	0
YEAR_OF_BIRTH	10251	60.56316457	64	57	2.232746333
COUNTRY_OF_BIRTH*	10251	2.934738074	3	1	0.24739418
SAMPLE_RACE*	10251	2.311091601	3	1	0.859093617
SAMPLE_SEX*	10251	1.489903424	2	1	0.499922433

C1DOB_Y*	10251	25.00975515	47	1	12.73684505
HAVING_HEALTHPLAN*	10251	3.788020681	4	1	0.413712394
FAMSIZE_	10251	3.099892693	15	1	1.654254231
TNFI_	10251	27048.72198	146942	-3	26287.04635
POVSTATUS_*	10251	1.964881475	3	1	0.532949211
REGION_*	10251	3.605404351	5	1	1.001991217
MARSTAT_KEY_*	10251	2.937664618	6	1	0.895907644
URBAN_RURAL_*	10251	3.776899815	4	1	0.442693516
JOBSNUM_	10251	7.652229051	43	-3	4.616061002
NUMCH_	10251	1.084772217	7	0	1.218048537
EVER_IN_POVERTY*	10251	1.669690762	2	1	0.470347346
WHEN_IN_POVERTY*	10251	2.162618281	4	1	1.353799183
INCOME_	10251	15215.97717	74283	-3	14427.96114
INCOME_MAX	10251	51172.51771	343830	0	56767.42853
EVER_EDU_LOAN*	5152	2.375776398	3	1	0.502468551
EVER_DIVORCED_*	10251	2.158618671	3	1	0.376127384
EVER_UNEMPLOYED_*	10251	1.587162228	2	1	0.49236815

EMP_STATUS_*	10251	1.44083504	4	1	0.833380799
AGE	10251	1929.436835	1933	1926	2.232746333
Black	10251	0.261828114	1	0	0.439651008

Table: 2 - Statistic for Female with birthplace in US.

	N	Mean	Max	Min	SD
ID	4901	5710.042644	12667	3	3406.273611
YEAR	4901	1990	1990	1990	0
YEAR_OF_BIRTH	4901	60.49316466	64	57	2.233649024
COUNTRY_OF_BIRTH*	4901	1	1	1	0
SAMPLE_RACE*	4901	2.334421547	3	1	0.868834909
SAMPLE_SEX*	4901	1	1	1	0
C1DOB_Y*	4901	19.44521526	40	1	11.12724676
HAVING_HEALTHPLAN*	4901	3.834931647	4	1	0.375651734
FAMSIZE_	4901	3.264639869	15	1	1.618500319
TNFI_	4901	26703.11243	146942	-3	25765.34932
POVSTATUS_*	4901	2.017139359	3	1	0.548943381
REGION_*	4901	3.589675576	5	1	0.971831595
MARSTAT_KEY_*	4901	2.034686799	5	1	0.925169951

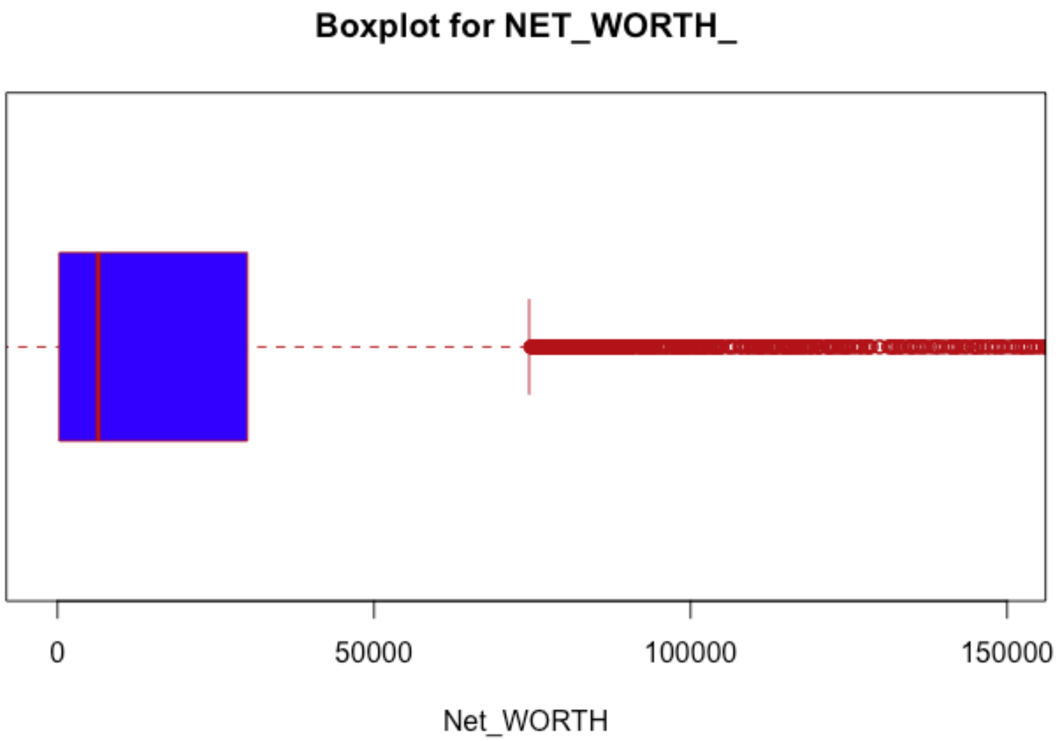
WKSUEMP_PCY_	4901	1.803509488	52	-3	6.215447833
URBAN_RURAL_*	4901	2.778004489	3	1	0.417589996
JOBSNUM_	4901	7.185268313	38	-3	4.362731226
NUMCH_	4901	1.352377066	7	0	1.244288569
AGE_1STCHILD*	4901	17.04550092	37	1	10.70060892
EVER_IN_POVERTY*	4901	1.694144052	2	1	0.460816032
WHEN_IN_POVERTY*	4901	2.078963477	4	1	1.336034426
INCOME_	4901	11233.92349	74283	-3	11713.82602
INCOME_MAX	4901	38663.57539	343830	0	39992.45627
EVER_EDU_LOAN*	2523	2.38208482	3	1	0.502835037
EDU_DEGREE*	4901	5.234033871	10	1	2.009495985
EVER_DIVORCED_*	4901	2.193225872	3	1	0.403053422
EVER_UNEMPLOYED_*	4901	1.577637217	2	1	0.493986086
HOURS_WORKED_PER_WEEK_	4901	8.612528055	138	-4	23.99668851
MAJOR_1_	4901	585.7867782	9996	-4	1058.748712
MAJOR_2_	4901	0.849418486	2104	-4	81.78743588
EDU_MAJOR	4901	671.979596	9996	-4	1160.401754

AMT_EDU_LOAN_	4901	94.25117323	75000	-4	1375.228647
TOTAL_EDU_LOAN	4901	2302.624362	99999	-5	7171.045721
CAL_YEAR_JOBS_	4901	1.714548051	7	0	1.041942135
NET_WORTH_	4901	38268.17017	841832	-890000	120339.5699
EMP_STATUS_*	4901	1.564374617	4	1	0.877954465
AGE	4901	1929.506835	1933	1926	2.233649024

From above table we can see the Sample sex- Female in Country of birth – IN THE US

2 Liner model

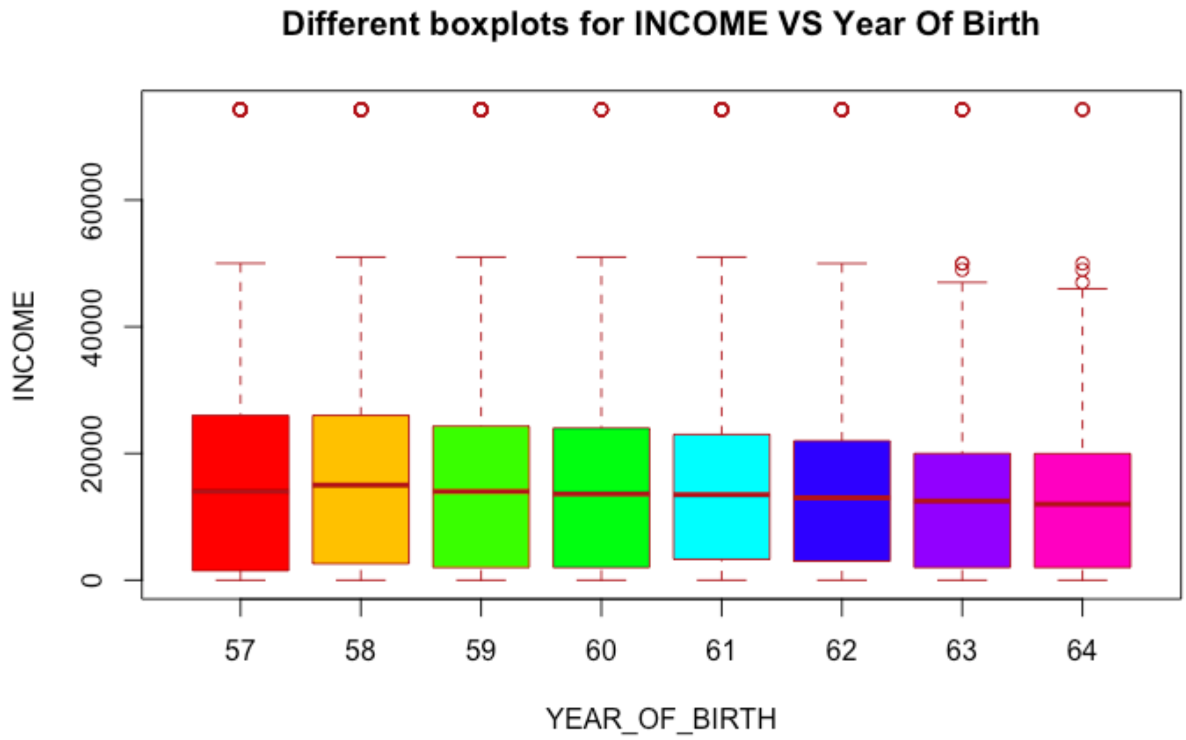
2.1 Boxplot for Net_Worth: The plot displayed below help determine to get Net_worth



Observation: Most of the values are <50,000 only few values are >50,000 which shown as an outlier in a plot.

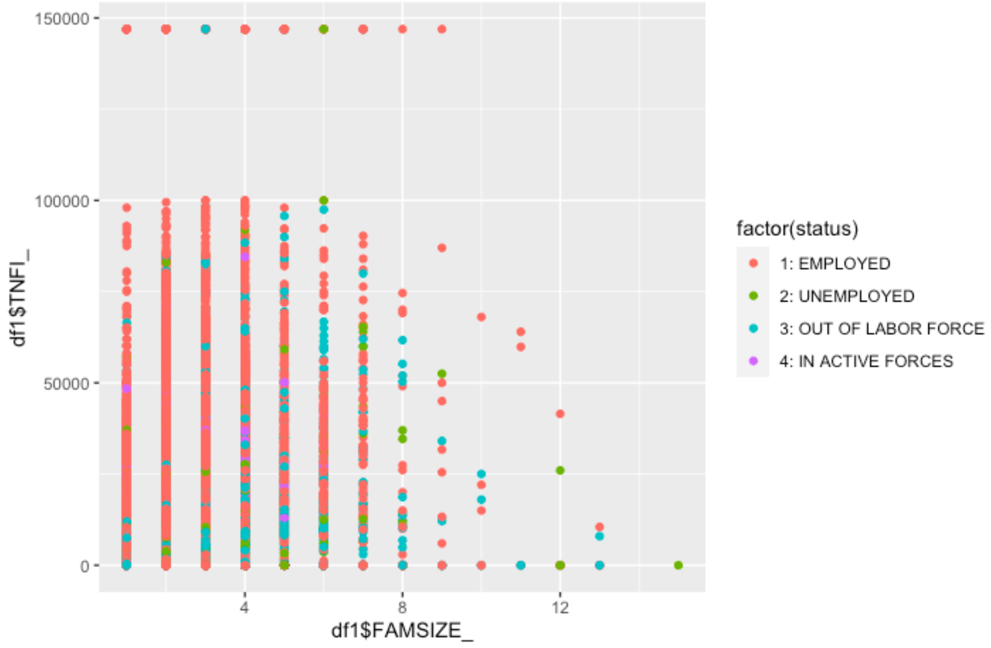
Formula: Q1-1.5*IQR - LOWER OUTLIER
Q3+1.5*IQR – HIGHER OUTLIER (Data point is a outlier more than 1.5 above third quantile below first quantile)

2.2 BOXPLOT- Diffeent Boxplots for Income Vs Year Of Birth



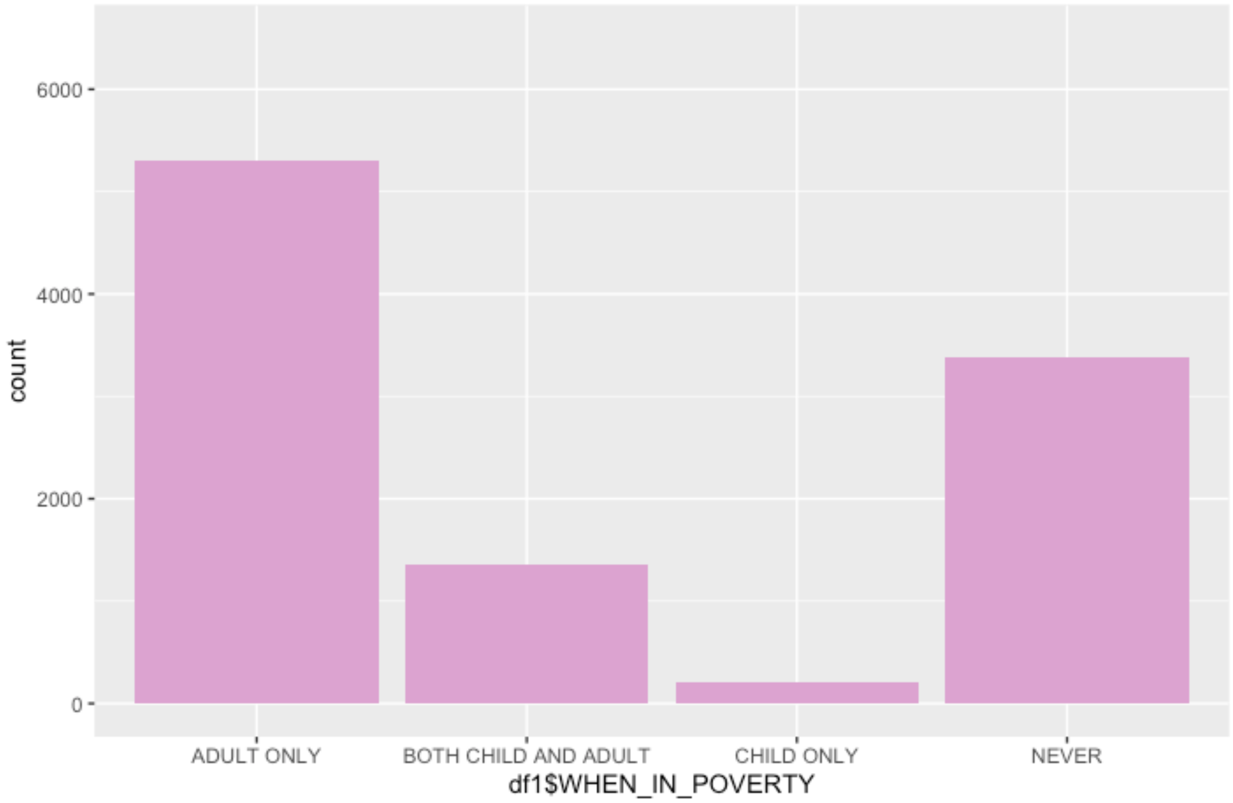
Observation: The plot displayed above shows that as year of the birth increases Income significantly decreases. More of the income varies between (0 – 30,000) and the average income for all year of birth is <20,000.

2.3 Scatterplots: Scatter plot for Employment Distribution



Observation: An observation we have found is that as Fam size increases Employment decreases we can clearly observe that Fam size > 6 has less employment. Here, TNFI is Total no.of Family.

2.4 BARPLOT: Barplot for When in Poverty



Observation: Count of samples by the time they were in poverty.
5300 people were in poverty only as an Adult, 3500 people were under poverty line both as a child and Adult, 400 people under poverty line as a Child only and 3200 people were never under poverty line.

Milestone: Project 2

Introduction: The data provided is a typical bookkeeping data set held by a population update based on Region by considering the sample sex, Race and getting the insights of the income, Income of the family members according to size.

Here In document, I consider few categorical variable (HAVING_HEALTHPLAN, MARSTAT_KEY, SAMPLE_RACE) and numerical variable (NET_WORTH_, INCOME_, YEAR_OF_BIRTH). These numerical and categorical data used describe the complete parameter and statistics of the given data.

After loading the dataset, we perform the cleaning step. It comprises of primarily removing data records which have irrelevant data (this could include incompatible data types, outliers, extreme values). For these values if found in dataset we replace them with the mean of the dataset.

The first section of analysis is descriptive statistics. Descriptive tables for dataset considering categorical variables, numerical variable and exploratory data analysis are shown followed by inferential analysis using histogram, boxplots and scatterplots. Datasets are easy to manipulate, model and visualize. It important to clean, messy dataset and perform the analysis (Wickham 2014).

ANALYSIS

1.DESCRIPTIVE STATISTICS

1.1 Statistic descriptive table:

Table 1: Summary stats of Dataset - Analyzed data using below data fields.

DATASET

	N	Mean	Max	Min	SD
ID	10251	5862.310409	12679	2	3463.713809
YEAR	10251	1990	1990	1990	0
YEAR_OF_BIRTH	10251	60.56316457	64	57	2.232746333
COUNTRY_OF_BIRTH*	10251	2.934738074	3	1	0.24739418
SAMPLE_RACE*	10251	2.311091601	3	1	0.859093617
SAMPLE_SEX*	10251	1.489903424	2	1	0.499922433
C1DOB_Y*	10251	25.00975515	47	1	12.73684505
HAVING_HEALTHPLAN*	10251	3.788020681	4	1	0.413712394
FAMSIZE_	10251	3.099892693	15	1	1.654254231
TNFI_	10251	27048.72198	146942	-3	26287.04635

POVSTATUS_*	10251	1.964881475	3	1	0.532949211
REGION_*	10251	3.605404351	5	1	1.001991217
MARSTAT_KEY_*	10251	2.937664618	6	1	0.895907644
URBAN_RURAL_*	10251	3.776899815	4	1	0.442693516
JOBSNUM_	10251	7.652229051	43	-3	4.616061002
NUMCH_	10251	1.084772217	7	0	1.218048537
EVER_IN_POVERTY*	10251	1.669690762	2	1	0.470347346
WHEN_IN_POVERTY*	10251	2.162618281	4	1	1.353799183
INCOME_	10251	15215.97717	74283	-3	14427.96114
INCOME_MAX	10251	51172.51771	343830	0	56767.42853
EVER_EDU_LOAN*	5152	2.375776398	3	1	0.502468551
EVER_DIVORCED_*	10251	2.158618671	3	1	0.376127384
EVER_UNEMPLOYED_*	10251	1.587162228	2	1	0.49236815
EMP_STATUS_*	10251	1.44083504	4	1	0.833380799
AGE	10251	1929.436835	1933	1926	2.232746333
Black	10251	0.261828114	1	0	0.439651008

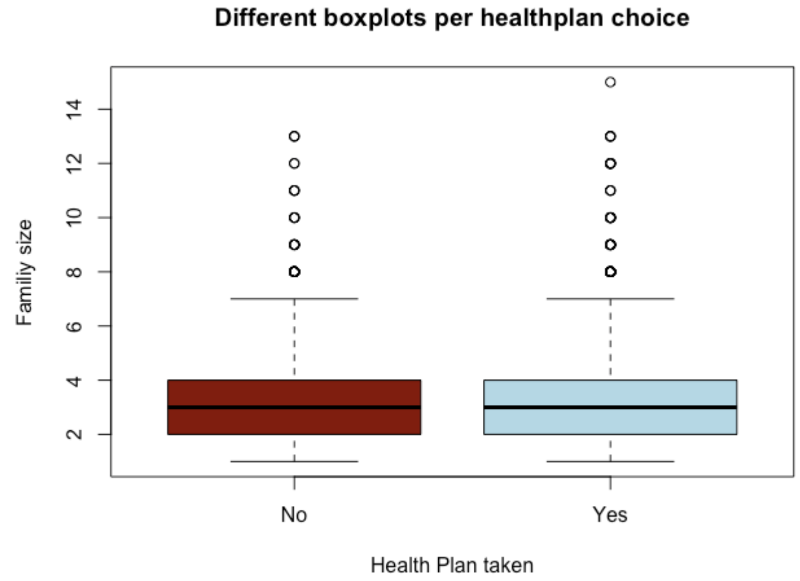
Data planning is not just a first step; it must be replicated several times during the analysis as new problems occur or new data is gathered. Conducted test statistics for the difference of the two-sample means. When independent sample are large Null hypothesis

Null Hypothesis: $H0$: $\mu 1-\mu 2=\mu 0$

Alternative Hypothesis: $H1$: $\mu 1-\mu 2\neq\mu 0$

When independent sample are less of the two-sample means $\mu 0$ = the hypothesized difference usually 0.

Question 1: Do people who have health plan or not have the same family?



Observation: mean of having health plan and family not having health plan are same. By conducting Two Sample test - Two sample test runs with only two mean result and two sample t-test compares mean of predetermined value to get the significant level (>, =, or <).

Table: 2 sample t-test result

T-value	df	P-value	95% CI	Mean of x.	Mean of y
-22.869	2975	< 2.2e-16	-0.1254170 -0.1254170	3.134842	3.17

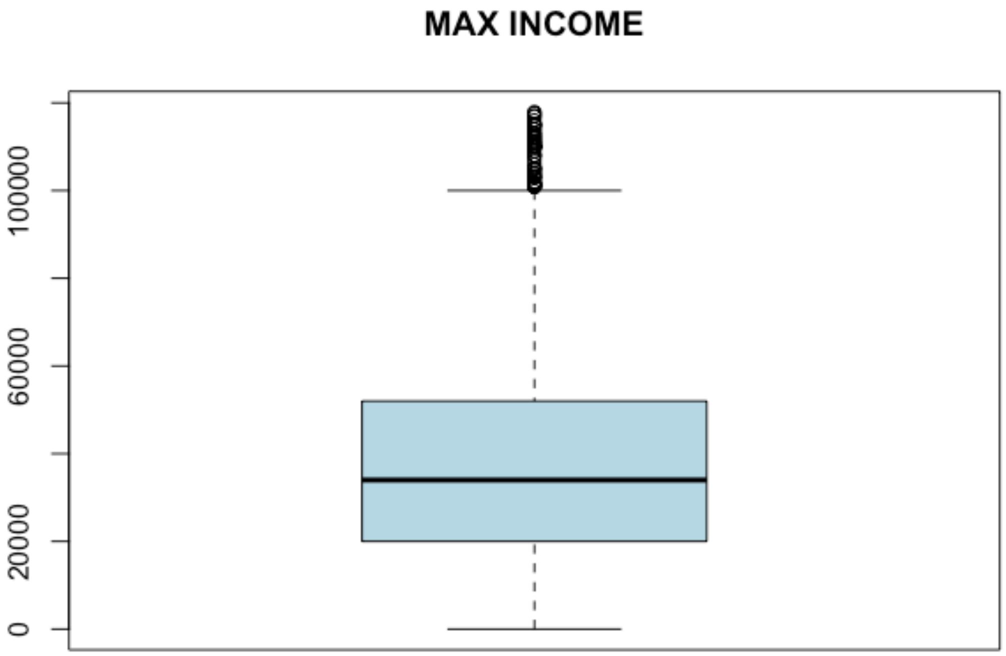
By comparing mean of having health plan(x) and mean of not having health plan(y) to see any significant difference in family size
We test if mean of x – mean of y = 3.13 - 3.17 = 0
Therefore, there is no difference between mean.
Here, P-value: < 2.2e-16 is less significant than alpha = 0.05 it rejects the null hypothesis (area based rejection)

Interpretation of the result:

p-value is 2.2e-16 is less significant level alpha = 0.05
In the above t.test statistic value t = -22.869
Degree of freedom df = 2975
P - value for alternative hypothesis is not equal to 1
Confidence interval of the mean at 95% (conf.int = (low: -0.1254170, high - 0.0523264)

sample estimates:
mean of having health plan(x) = 3.134842
mean of not having health plan(y) = 3.171387

Question 2: The maximum income of 30% population is more than 10000.



Analysis: By conducting one Sample test – One sample test runs with only one mean result and one sample t-test compares mean of predetermined value to get the significant level (>, =, or <).

Table: 2 sample t-test result

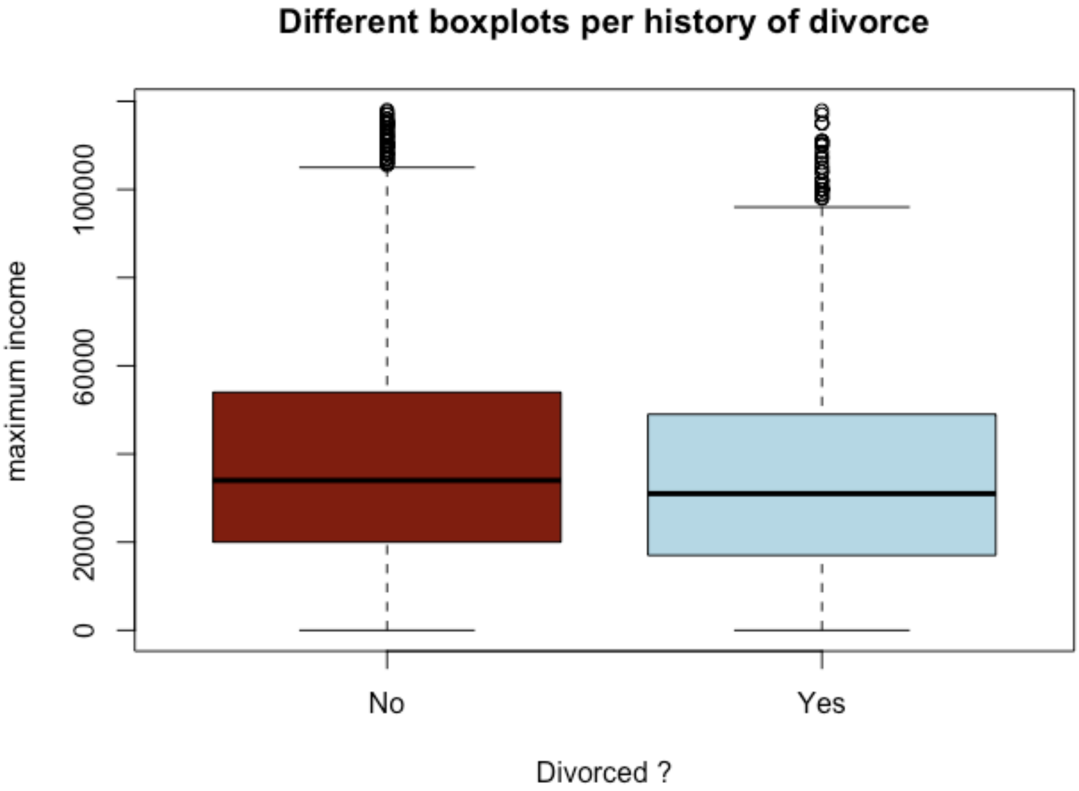
T-value	df	P-value	95% CI	Mean of x
101.08	4177	< 2.2e-16	37353.08 38830.75	38091.92

p-value is 2.2e-16 is less significant level alpha = 0.05
Since alpha value is less than alpha value it rejects the null hypothesis.

Interpretation of the result:
p-value is 2.2e-16 is less significant level alpha = 0.05
In the above t.test statistic value t = 101.08

Degree of freedom df = 4177
P - value for alternative hypothesis is not equal to 1
Confidence interval of the mean at 95% (conf.int = (low: 37353.08, high - 38830.75)
sample estimates:
mean of Max_ Income(x)   = 38091.92

Question 3: Do people who have or have not been ever divorced have different maximum income.



Inferential Statistics: Based on the boxplot we can say that the maximum income for people who has never been divorced is 50000.
By conducting Two Sample test - Two sample test runs with only two mean result and two sample t-test compares mean of predetermined value to get the significant level (>, =, or <).

Table: 2 sample t-test result

T-value	df	P-value	95% CI	Mean of x.	Mean of y

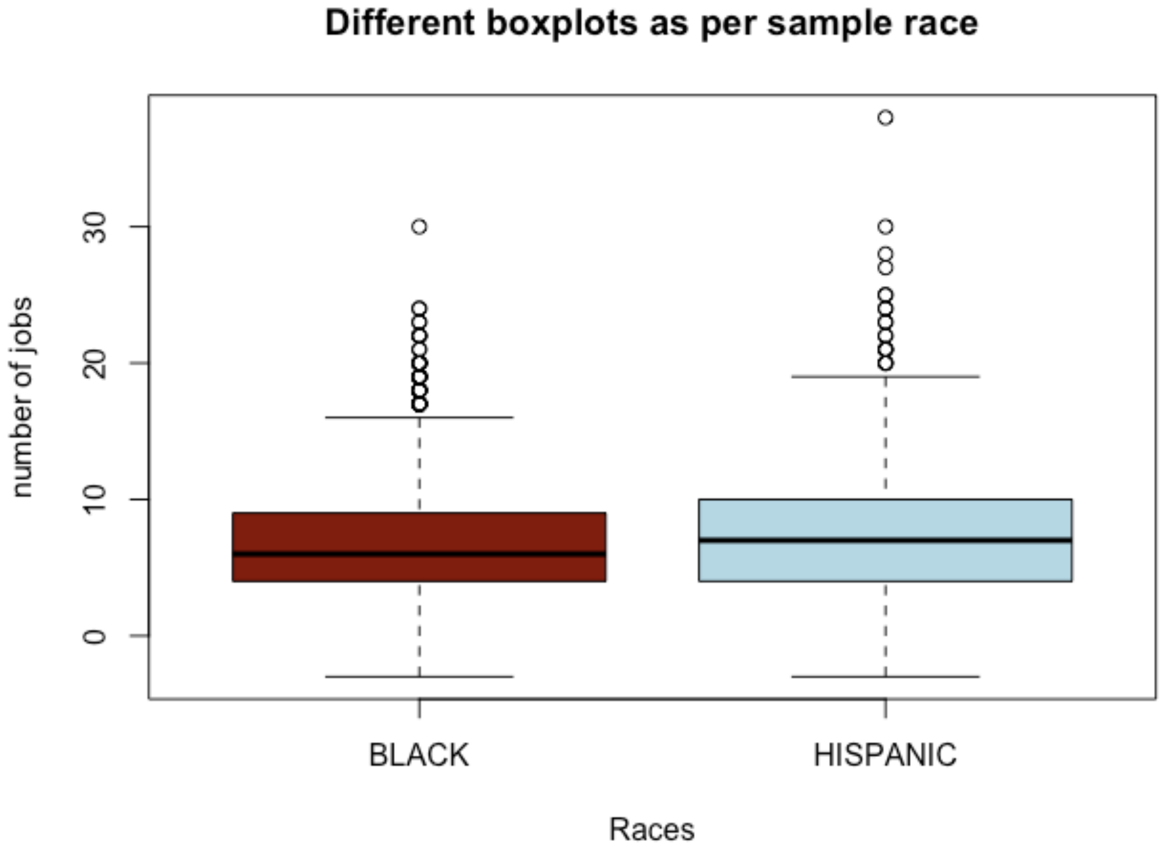
-3.8585	2351.8	0.0001172	-3984.96 -1298.68	36105.39	38747.22

By comparing mean of people who were ever divorced and mean of people who were never divorced(y) to see any significant difference in Maximum income. We test if mean of x – mean of y = 38747.22- 36105.39 = 2641.83 Therefore, there is no difference between mean. Here, P-value: 0.0001172 is less significant than alpha = 0.05 it rejects the null hypothesis (area based rejection)

Interpretation of the result:

p-value is 2.2e-16 is less significant level alpha = 0.05
In the above t.test statistic value t = -3.8585
Degree of freedom df = 2351.8
P - value for alternative hypothesis is not equal to 1
Confidence interval of the mean at 95% (conf.int = (low: -3984.96, high: -1298.68)
sample estimates:
mean of people who were ever divorced (x) = 36105.39
mean of people who were never divorced (y) = 38747.22

Question 4: Does the number of jobs differ in URBAN or Rural area?



Inferential Statistics: Based on data we have we can infer that the maximum number of jobs for people belonging to Hispanic Race is more than people belonging to African American Race. By conducting Two Sample test - Two sample test runs with only two mean result and two sample t-test compares mean of predetermined value to get the significant level (>, =, or <).

Table: 2 sample t-test result

T-value	df	P-value	95% CI	Mean of x.	Mean of y
-9.8654	3102.2	2.2e-16	-0.6782587 -0.1217594	6.851409	7.251418

By comparing mean of people with sample race Black(x) and mean of people with sample race Hispanic(y) to see any significant difference in number of jobs.
We test if mean of x – mean of y = 38747.22- 36105.39 = 2641.83
Therefore, there is no difference between mean.
Here, P-value: 0.0001172 is less significant than alpha = 0.05 it rejects the null hypothesis (area based rejection)

Interpretation of the result:
p-value is 2.2e-16 is less significant level alpha = 0.05
In the above t.test statistic value t = -9.8654

Degree of freedom df = 3102.2
P - value for alternative hypothesis is not equal to 1
Confidence interval of the mean at 95% (conf.int = (low: -0.6782587, high: -0.6782587)
sample estimates:
mean of people with sample race Black(x) = 6.851409
mean of people with sample race Hispanic(y) = 7.251418

FINAL PROJECT

Introduction: By analysing the complete dataset and by considering milestone 1 and milestone 2 my observation on few variables is analysed in this report. Complete insights from above reports. In first section we will explore to few variable from milestone 1 and analyses the data. In second section will conduct hypothesis testing on different variables and get null and alternative hypothesis testing.

Questions :

Question 1 : What is max net worth of 30% population is more than 35000?

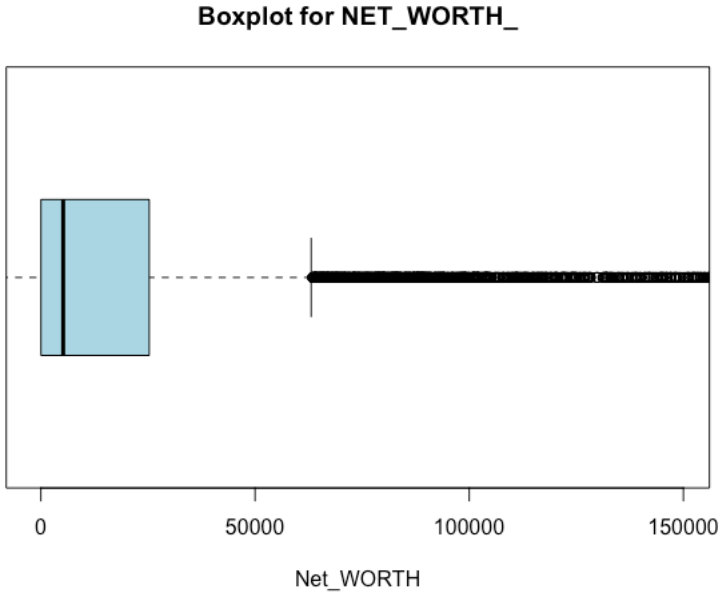
Question 2 : How year of birth is related to income?

Question 3 : How is family size proportional to employment?

Analyses:

Question 1 : What is max net worth of 30% population is more than 35000?

Result:



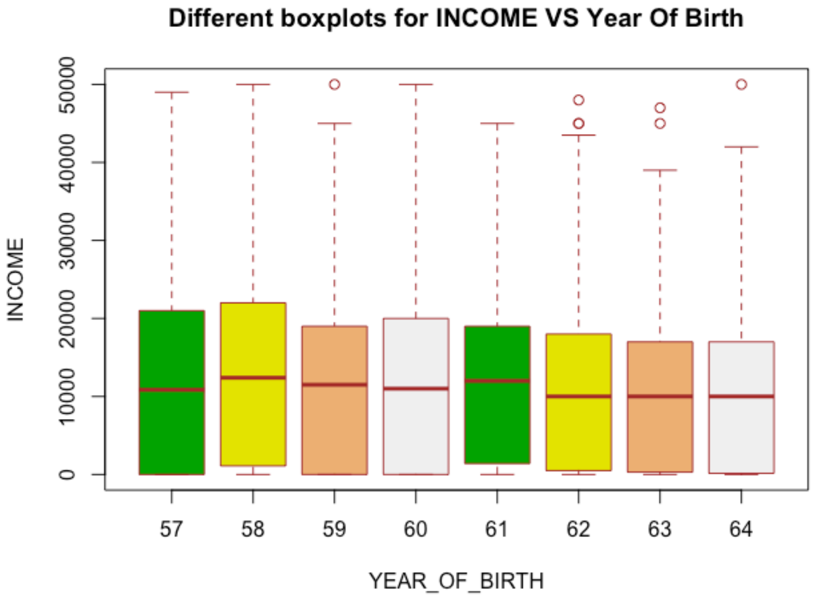
Analysis: By conducting one Sample test – One sample test runs with only one mean result and one sample t-test compares mean of predetermined value to get the significant level (>, =, or <).

Table: 2 sample t-test result

T-value	df	P-value	95% CI	Mean of x
29.277	9551	< 2.2e-16	30143.22 34469.33	32306.28

p-value is 2.2e-16 is less significant level alpha = 0.05
Since alpha value is less than alpha value it rejects the null hypothesis.

Question 2 : How year of birth is related to income?



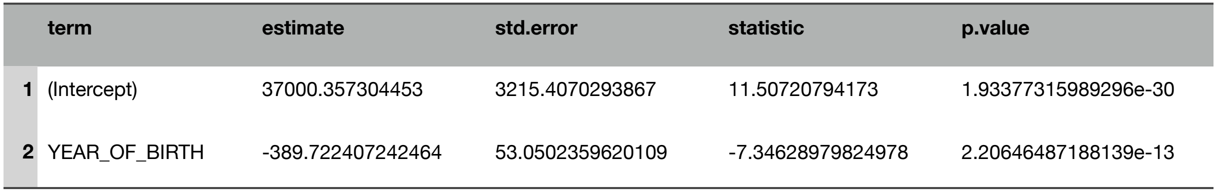
Inferential Statistics: The plot displayed above shows that as year of the birth increases Income significantly decreases. More of the income varies between (0 – 30,000) and the average income for all year of birth is <20,000, by this we can clear predict that young age people are getting more income.
By conducting Two Sample test - Two sample test runs with only two mean result and two sample t-test compares mean of predetermined value to get the significant level (>, =, or <).

Table: 2 sample t-test result

T-value	df	P-value	95% CI	Mean of x.	Mean of y

70.251	4177	2.2e-16	11398.95 12053.39	11786.83485	60.66587

Regression table1



	term	estimate	std.error	statistic	p.value
1	(Intercept)	37000.357304453	3215.4070293867	11.50720794173	1.93377315989296e-30
2	YEAR_OF_BIRTH	-389.722407242464	53.0502359620109	-7.34628979824978	2.20646487188139e-13

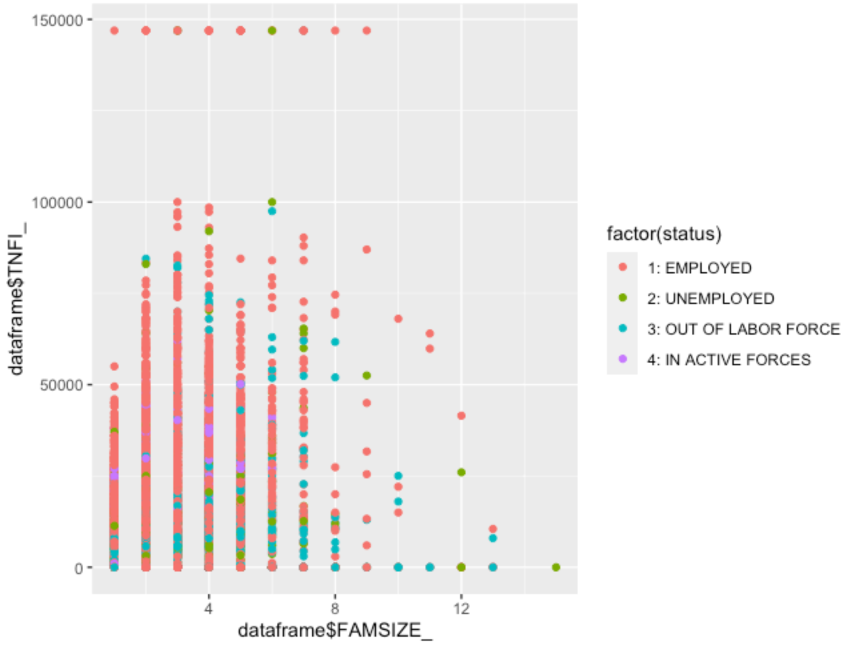
In Regression table we can see that there is a positive correlation between Year of birth and Income. Higher the year of birth, less is the income.

By comparing mean of people with sample race Black(x) and mean of people with sample race Hispanic(y) to see any significant difference in number of jobs.
We test if mean of x – mean of y = 11786.83- 60.66587 = 11726.1
Therefore, there is no difference between mean.
Here, P-value: 2.2e-16 is less significant than alpha = 0.05 it rejects the null hypothesis (area based rejection)

Interpretation of the result:
p-value is 2.2e-16 is less significant level alpha = 0.05
In the above t-test statistic value t = 70.251

Degree of freedom df = 4177
P - value for alternative hypothesis is not equal to 1
Confidence interval of the mean at 95% (conf.int = (low: -0.6782587, high: -0.6782587)
sample estimates:
mean of Income = 11786.83485
mean of Year of birth = 60.66587
Intercept = 3700.357

Question 3 : How is family size proportional to employment?
Analysing employment distribution per family size and the total net Family income.



Scatter Plot for FAMSIZE vs TNFI

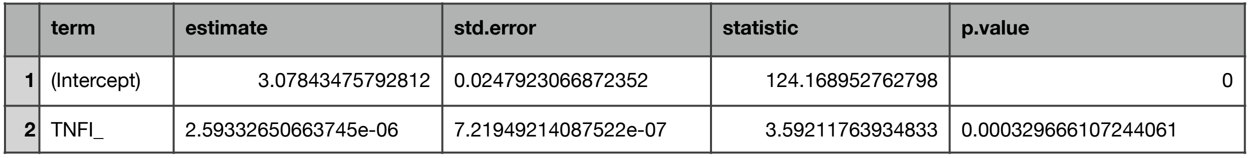
Inferential Statistics: An observation we have found is that as Fam size increases Employment decreases, we can clearly observe that Fam size > 6 has less employment. Here, TNFI is Total no. of Family. We can predict that to increase employment Fam size should be less than 4.

By conducting Two Sample test - Two sample test runs with only two mean result and two sample t-test compares mean of predetermined value to get the significant level (>, =, or <).

Table: 2 sample t-test result

T-value	**df**	**P-value**	**95% CI**	**Mean of x.**	**Mean of y**
-58.633	4177	2.2e-16	-20356.35 -19039.01	3.41551	19701.09813

Regression table 2



	term	estimate	std.error	statistic	p.value
1	(Intercept)	3.07843475792812	0.0247923066872352	124.168952762798	0
2	TNFI_	2.59332650663745e-06	7.21949214087522e-07	3.59211763934833	0.000329666107244061

In Regression table we can see that there is a positive correlation between family size and total number family . Higher the family size, less is the employment.

By comparing mean of people with sample race Black(x) and mean of people with sample race Hispanic(y) to see any significant difference in number of jobs.
We test if mean of x – mean of y = 11786.83- 60.66587 = 11726.1
Therefore, there is no difference between mean.
Here, P-value: 2.2e-16 is less significant than alpha = 0.05 it rejects the null hypothesis (area based rejection)

Interpretation of the result:
p-value is 2.2e-16 is less significant level alpha = 0.05
In the above t-test statistic value t = -58.633

Degree of freedom df = 4177
P - value for alternative hypothesis is not equal to 1
Confidence interval of the mean at 95% (conf.int = (low: -20356.35, high: -19039.01)
sample estimates:
mean of FAMSIZE_ = 3.41551
mean of TNFI = 19701.09813
Mean y – mean x = 19697.67

Intercept = 3.078

Conclusion: By analyzing the dataset, there are so many important insights. From descriptive Statistics table data using categorical variables and numerical variables. On exploratory data analysis reveals that there are more people were in people only as an Adult, Employment rate is decreasing on increasing the Family size and Income decreases on increasing the Age. Linear models used for prediction and linear models without removing outliner we can clearly observe the Net worth and Income.

After analyzing milestone 1 dataset and using various inferences, asking questions, and using hypothesis to draw conclusions and answer the questions it can be clearly said that hypothesis testing methods help us analyze data and answer important questions in order to make better sense of the data at hand and find hidden relationships among various parameters of the dataset and provides a framework for making determination related to the population.

References:

1.Allan B (2015)- Elementary statistics, edition 7). Retrieved January 19th, 2018, from https://bmalone.weebly.com/uploads/2/2/3/9/22391186/bluman_statistics_book.pdf/