**CANCER MORTALITY RATES PREDICTION**

By:

**Shivaramteja Keerthi**

Submitted to:

**Dr. Duoduo Liao**,

George Mason University,

AIT-580 : Analytics : Big Data to Information

Date: 12/13/19

**CANCER MORTALITY RATES PREDICTION**

**Abstract**

Cancer is a deadly disease. Though there is treatment for this, many people succumb to death. Data analytics is playing a prominent role in health care these days. This project is an attempt to explore how data analytics can be used to handle cancer related data. This project cancer mortality rates prediction is mainly determined to predict cancer death rates in different counties in the united states. The data for this project has been obtained from data.world website. This data is an aggregation of data from different websites like census.gov, cancer.gov and clinical trials.gov. Different contributors contributed the data and this has been aggregated resulting in a dataset with 3047 observations and 34 attributes. These attributes are mainly different factors contributing to cancer death rates in that county. The tasks involved in this project are: The first step is Data preprocessing which involves cleaning of data by removing missing values and handling the outliers. The second step is exploratory analysis which is exploring the patterns in data by plotting the necessary graphs which also includes correlation analysis. The third step is to perform data analysis using a model. Here in this project the data analysis task is to predict the cancer death rates. The model used is Linear Regression (Ordinary Least Squares Regression model). The fourth step is to perform hypothesis test to check our results by considering both null and alternative hypothesis. The data does not consist of external attributes like smoking, drinking but consists of status of that county like percent of poverty, percent of educated individuals etc.,

**Introduction**

This project is about predicting cancer death rates in different counties in the United States. The data mainly comprises of different factors contributing to cancer deaths in that county. The main tasks includes data cleaning, data exploration, extracting significant features and using them to predict death rates and performing hypothesis test. Data exploration includes plotting of scatter plots, box plots and other necessary plots for understanding the data.

This data is a result of aggregation of different data sets from census.gov, cancer.gov and clinical trails.gov. This data is collected from data.world website.

**Objectives**

The main objectives of this project are:

* Predicting the death rate in different counties in the United States
* Comparing the observed and the predicted death rate using hypothesis test

**The Dataset:**

**Dataset selection:**

The dataset has been taken from data.world website. The source is

[**https://data.world/nrippner/ols-regression-challenge**](https://data.world/nrippner/ols-regression-challenge)

**Dataset Description:**

This data set consists of 3047 observations and 34 attributes. All the attributes are numerical except the geography and binnedInc attributes. The size of the data set is 717 KB.

**Who collected the data?:**

This data has been obtained from data.world website. Data.world is the world’s largest collaborative data company, which is free and open to public. In data.world people discover data, share analysis, and team up on everything. data.world is the modern catalog for data and analysis. The data catalog unites and classifies your data, metadata, and analysis—no matter where it lives. The modern, intuitive user experience brings together employees of all roles, backgrounds, and skills to collaborate using the tools they already love. And the knowledge graph keeps data connected to everything people need to find, understand, and use it. As a result, your data, analysis, and expertise become more discoverable, trustworthy, and reusable**.**

**Need:**

As already mentioned, data.world is a collaborative data company it collected data from various fields like health, sports and finance etc., . This data comes under the health field. This data has been placed in the website and a challenge was introduced to predict the death rate. The contributors shared their data from different websites and all of them have been aggregated into one dataset.

**The Potential questions that could be answered are:**

1. Is there a model to predict cancer death rate?
2. Of all the 34 attributes what are the significant predictors?
3. How to handle missing values?

**Any issues:**

There are no privacy issues with this dataset as data.world is open and free to public. The data quality is good except for one attribute with a lot of missing values.

**Dataset Schema**

The attributes in the data set are described as follows:

**TARGET\_deathRate:** Dependent variable. Mean *per capita* (100,000) cancer mortalities

**avgAnnCount:** Mean number of reported cases of cancer diagnosed annually **avgDeathsPerYear:** Mean number of reported mortalities due to cance

**incidenceRate:** Mean *per capita* (100,000) cancer diagoses

**medianIncome:** Median income per county

**popEst2015:** Population of county

**povertyPercent:** Percent of populace in poverty

**studyPerCap:** *Per capita* number of cancer-related clinical trials per county

**binnedInc:** Median income per capita binned by decile

**MedianAge:** Median age of county residents

**MedianAgeMale:** Median age of male county residents

**MedianAgeFemale:** Median age of female county residents

**Geography:** County name

**AvgHouseholdSize:** Mean household size of county

**PercentMarried:** Percent of county residents who are married

**PctNoHS18\_24:** Percent of county residents ages 18-24 highest education attained: less than high school

**PctHS18\_24:** Percent of county residents ages 18-24 highest education attained: high school diploma

**PctSomeCol18\_24:** Percent of county residents ages 18-24 highest education attained: some college

**PctBachDeg18\_24:** Percent of county residents ages 18-24 highest education attained: bachelor's degree

**PctHS25\_Over:** Percent of county residents ages 25 and over highest education attained: high school diploma

**PctBachDeg25\_Over:** Percent of county residents ages 25 and over highest education attained: bachelor's degree

**PctEmployed16\_Over:** Percent of county residents ages 16 and over employed

**PctUnemployed16\_Over:** Percent of county residents ages 16 and over unemployed

**PctPrivateCoverage:** Percent of county residents with private health coverage

**PctPrivateCoverageAlone:** Percent of county residents with private health coverage alone (no public assistance)

**PctEmpPrivCoverage:** Percent of county residents with employee-provided private health coverage

**PctPublicCoverage:** Percent of county residents with government-provided health coverage

**PctPubliceCoverageAlone:** Percent of county residents with government-provided health coverage alone

**PctWhite:** Percent of county residents who identify as White

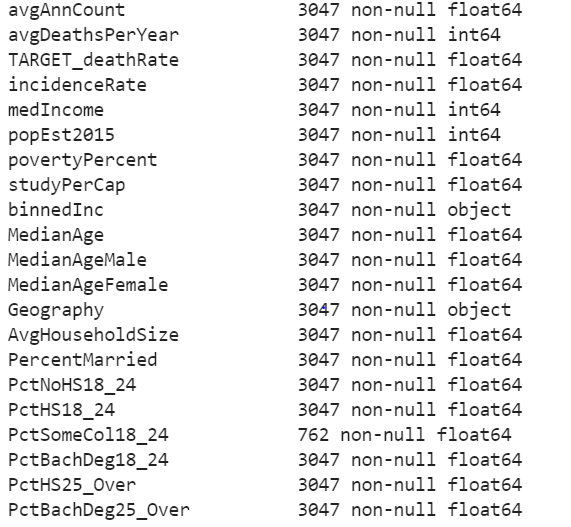
**PctBlack:** Percent of county residents who identify as Black

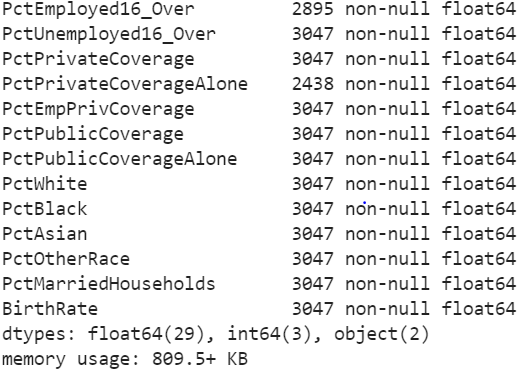
**PctAsian:** Percent of county residents who identify as Asian

**PctOtherRace:** Percent of county residents who identify in a category which is not White, Black, or Asian

**PctMarriedHouseholds:** Percent of married households

**BirthRate:** Number of live births relative to number of women in county



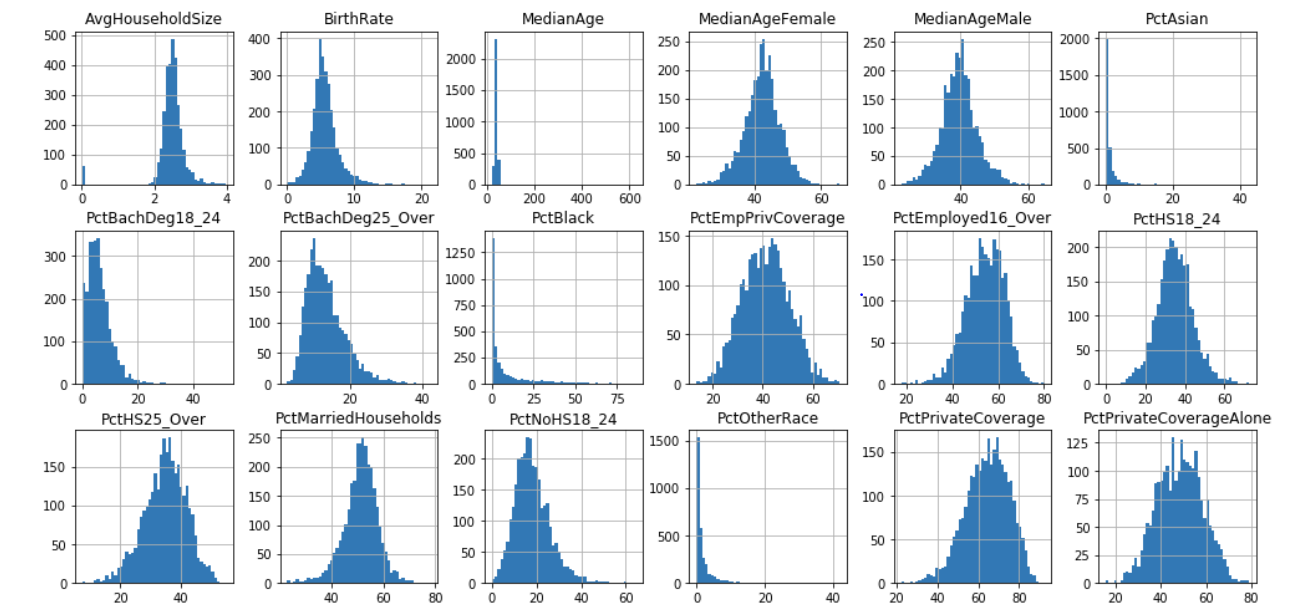


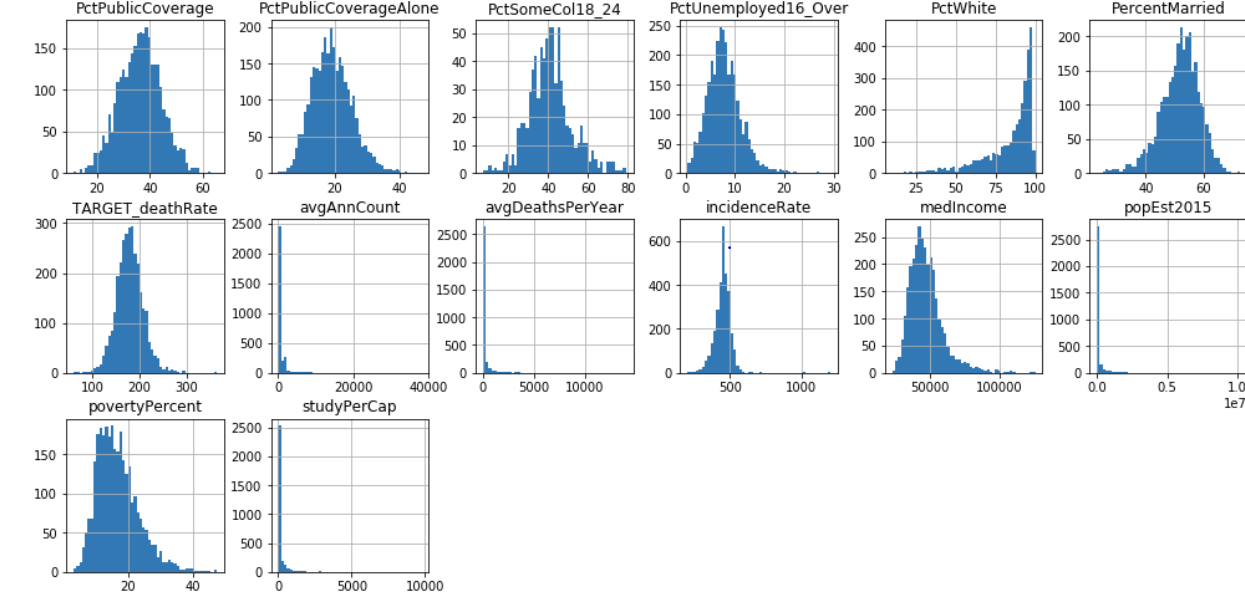
Screen shots for dataset description

**Dataset preprocessing**

**Exploring the data**

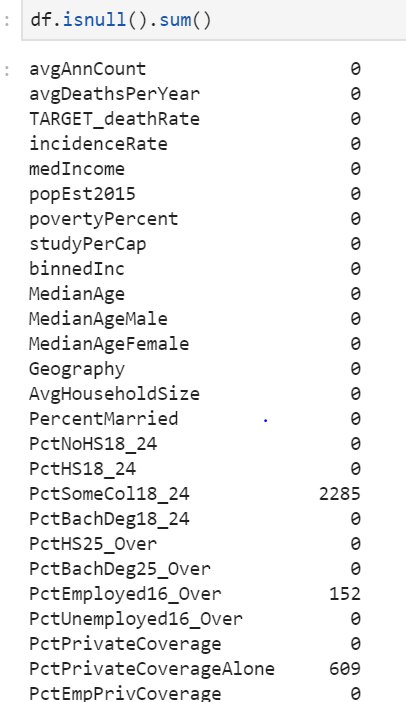
Before modelling there is a need to explore the data to understand how data is distributed. Histograms are plotted to describe the distribution of data.

****

****

All these histograms are unimodal, some of them are left skewed and some are right skewed. Right skewed histograms indicate there are outliers in the attribute, and they are greater than the mode. Left skewed indicate outliers less than the mode. Attributes with percentages are same alike because they are in the same range.

**Checking the missing values:**

****

The above screenshot describes there are missing values in three attributes: PctSomecol18\_24, PctEmployed16\_Over, PctPrivateCoverageAlone.

**Understanding attributes with missing values:**

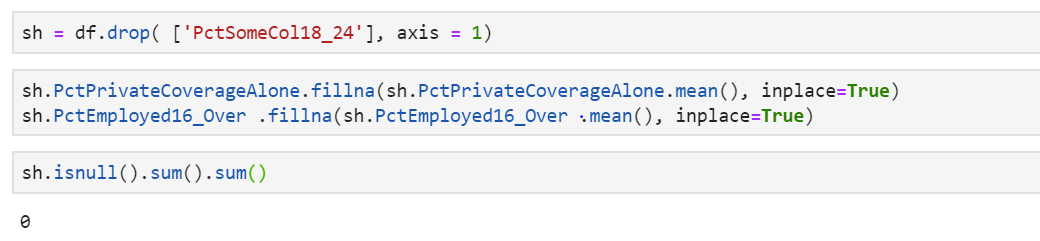
**A picture containing clock

Description automatically generated**

It is observed PctSomeCol18\_24 has many missing values and also many outliers, so it has to be deleted. The other two attributes are comparatively good so data imputation is performed on them.

**Data Cleaning:**

The attributePctSomeCol18\_24 has been dropped and the remaining two attributes missing values have been imputed by the mean values.

****

The above screenshot shows that after dropping PctSomeCol18\_24 and imputing mean values in other two attributes the number of missing values is now zero.

**System Architecture**

The proposed system consists of the following steps:

The first step is data preprocessing which includes handling missing values and outliers. The second step is exploring the data by plotting necessary plots like scatter plot, box plot and finding correlation between the attributes and eliminating the weakly correlated attributes. Third step is to perform analysis using any classification/prediction algorithm here in this case predicting death rate using linear regression technique and examining results. Fourth step is performing tests.

Exploratory analysis (plotting various plots and understanding data using correlation)

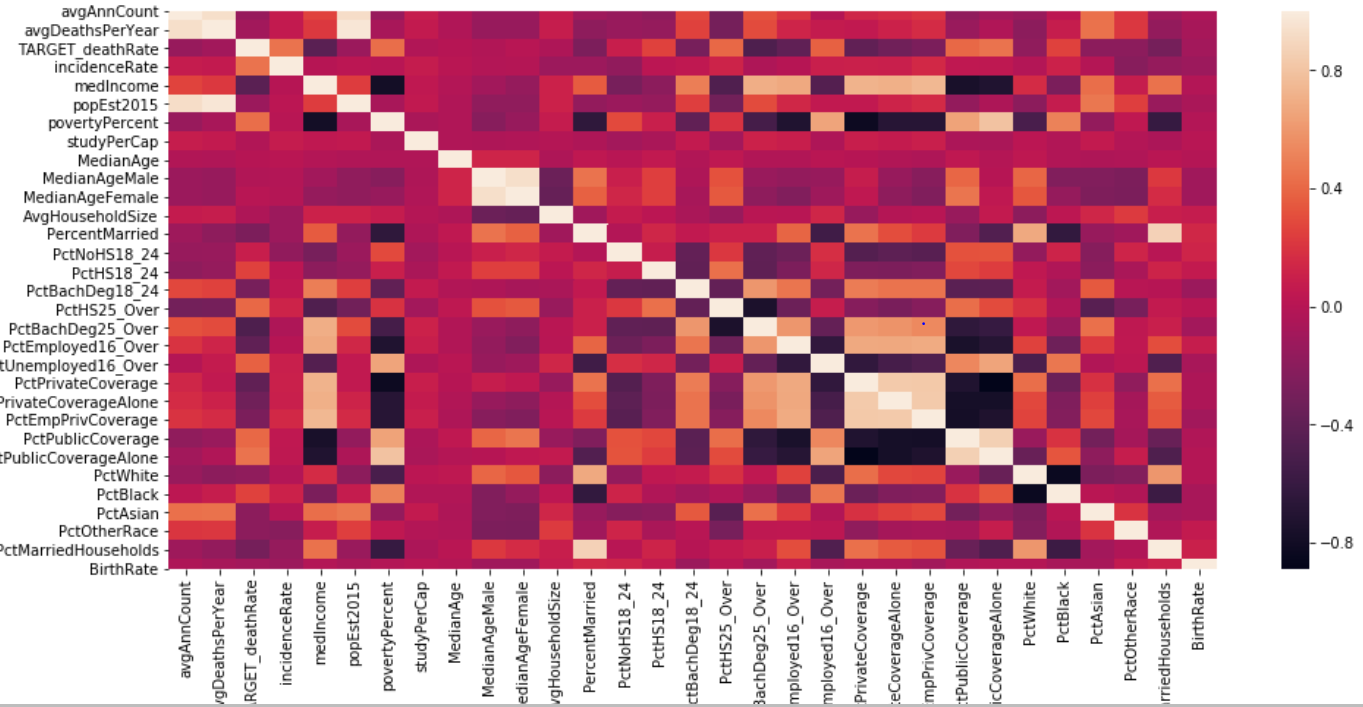
Performing tests (hypothesis test)

Performing data analytics and examining results

Data pre- processing (handling missing values)

**Data Processing:**

1. **Correlation analysis:**

****

The above plot is a heat map describing all the correlations. From the above picture it can be inferred that out of 31 attributes( one dropped column and two object attributes from 34) only 7 attributes are moderately correlated with threshold values ranging from 0.4 to 0.5.

**Understanding the correlated attributes:**

**A close up of a map

Description automatically generated**

**A close up of a map

Description automatically generated**

The above screenshots are scatterplots for the correlated attributes with respect to the target variable. It is observed that only povertypercent and pctpubliccoveragealone attributes are more linearly distributed than the other attributes.

**Data Analytics Algorithm**

The next step after understanding the correlated attributes is to predict the target variable using the seven correlated attributes. Here the target variable is TARGET\_deathRate and the predictors are incidenceRate, medIncome, povertyPercent, PctHS25\_over, PctBachDeg25\_over, PctPublicCoverage, PctPublicCoverageAlone. Since there are a number of predictors and only one response variable the algorithm used will be Linear regression.

**Linear Regression**

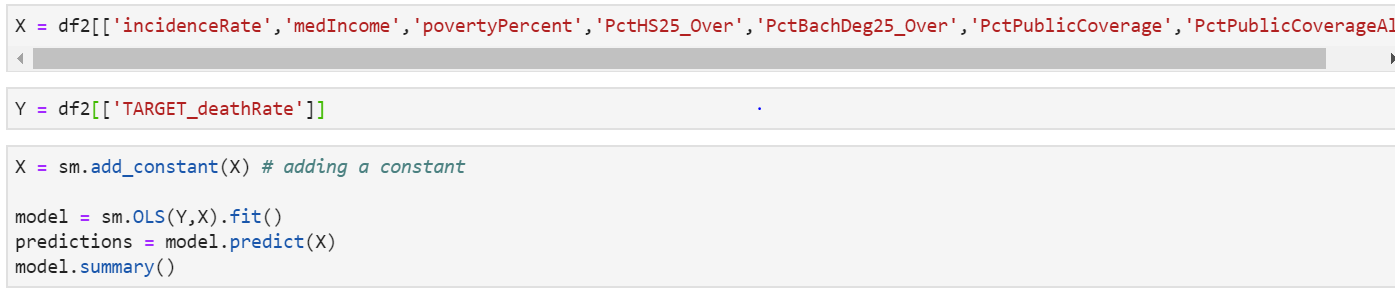
Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

Before attempting to fit a linear model to observed data, a modeler should first determine whether or not there is a relationship between the variables of interest. This does not necessarily imply that one variable causes the other (for example, higher SAT scores do not cause higher college grades), but that there is some significant association between the two variables. A scatterplot can be a helpful tool in determining the strength of the relationship between two variables. If there appears to be no association between the proposed explanatory and dependent variables (i.e., the scatterplot does not indicate any increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model. A valuable numerical measure of association between two variables is the correlation coefficient, which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables.

A linear regression line has an equation of the form Y = a + bX, where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, and a is the intercept (the value of y when x = 0).

**Using Linear Regression:**

Here there are seven explanatory variables and one dependent variable that is TARGET\_deathRate. The seven explonatory variables and one dependent variable are fitted into the model and results are observed.



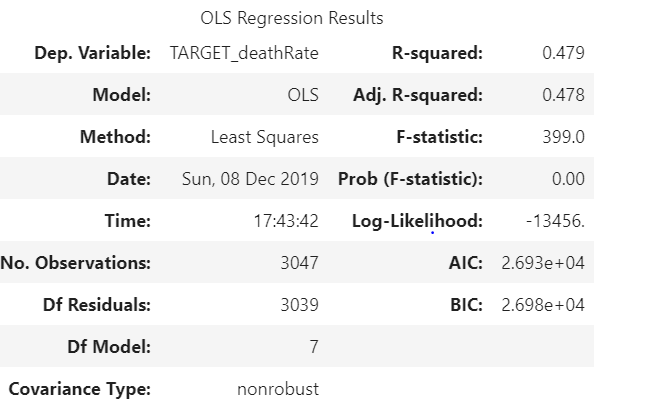
The predictors are arranged in a data frame X and the response variable is arranged in a data frame Y. Both X and Y are fitted into the regression model.

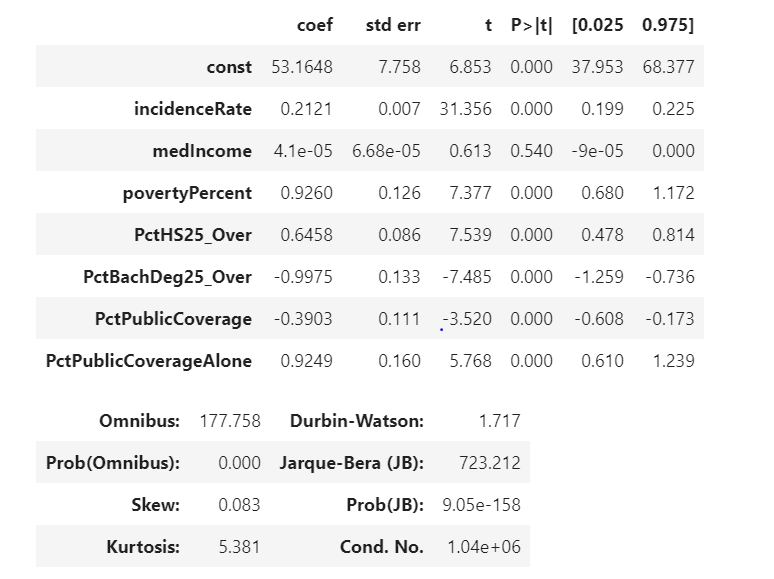
**Software Used:**

The software platforms used are **Python** and **Tableau.** Python for data preprocessing, data analysis and performing tests. Tableau for data visualizations.

**Experimental Results and Analysis**

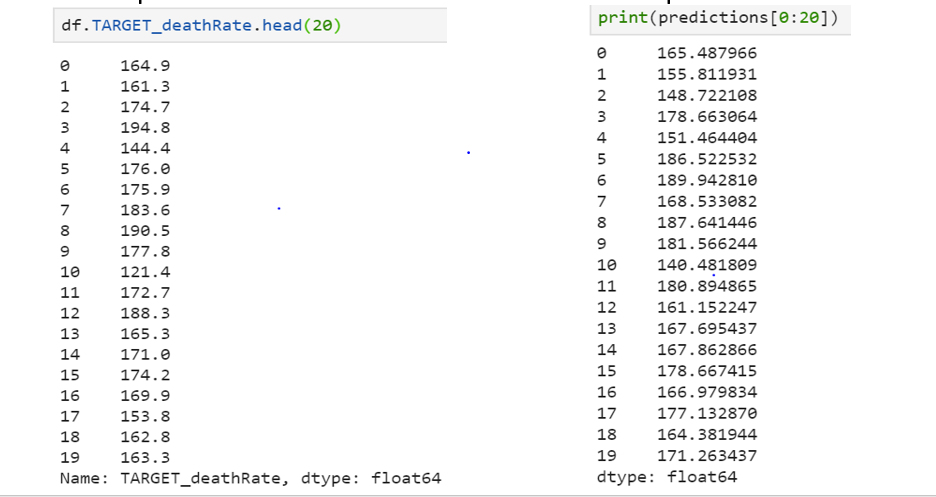
The summary of the model is:



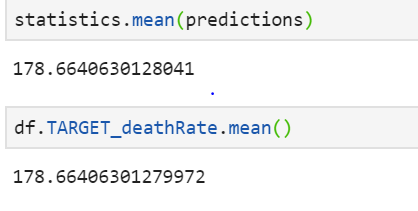
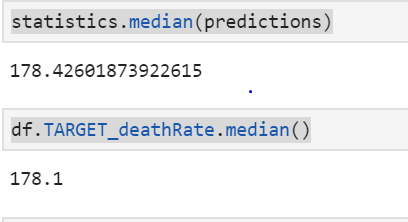


The above screen shots show that the R2 value for the model is 0.47 and the significant predictors according to the p-values are incidenceRate, povertyPercent, PctHS25\_over, PctBachDeg25\_over, PctPublicCoverage, PctPublicCoverageAlone.

Comparison between observed and predicted death rates are as follows:

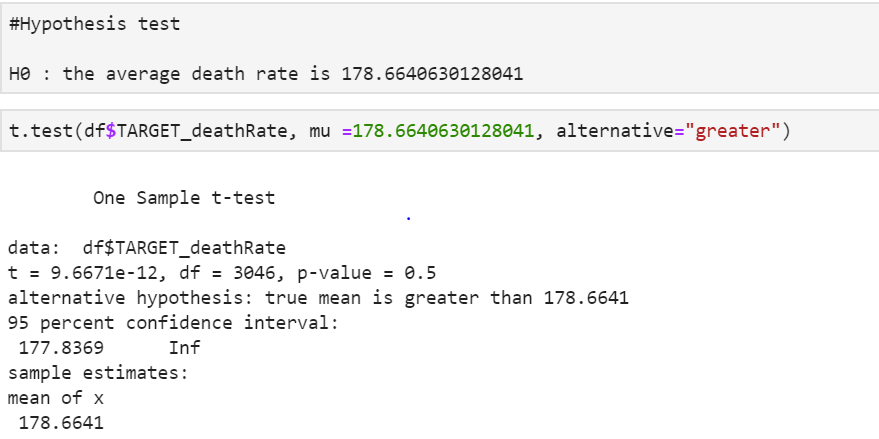


Here it can be inferred that some predicted values are comparatively very near to the observed values.

The mean of both observed and predicted death rates are almost the same. Since average is not a good metric to draw conclusions the median values are also observed. The medians are very close to each other. Thus it can be inferred that the regression model is significantly accurate. Thus the two potential questions have been answered: The linear regression model is used to predict the cancer death rates and the significant predictors are are incidenceRate, povertyPercent, PctHS25\_over, PctBachDeg25\_over, PctPublicCoverage, PctPublicCoverageAlone.

**Performing Hypothesis test**

****

Here a hypothesis test is performed to check whether the average death rate predicted is correct. Here the p-value obtained is 0.5 which is greater than 0.05 which can be concluded that it is **failed to reject** Null hypothesis.

**Conclusions and Future Scope**

* The Ordinary Least Square(OLS) regression model is used to predict cancer death rates for different counties. The adjusted R2 value is 0.478 which says our model is good enough for the data to be linearly fitted.
* More visualizations can be added if zip codes of the counties can be included in the dataset
* Since this data is about cancer, external variables like smoking, drinking and environment contamination can also be added and correlations can be checked.

**Lessons learnt**

* Handling missing values
* How to apply correct set of algorithms

**References**

1. Rippner, Noah. “OLS Regression Challenge”. data.world/nrippner/ols-regression-challenge. Accessed on December 13, 2019.
2. Stat.yale.edu.”Linear regression”.stat.yale.edu/Courses/1997-98/101/linreg.htm. Accessed on December 13, 2019.
3. Bronshtein, Adi.” Simple and Multiple Linear Regression in Python”. *Quick introduction to linear regression in python.* towardsdatascience.com/simple-and-multiple-linear-regression-in-python-c928425168f9. Accessed on December 13, 2019.

**What I have learnt in this course?**

This course was very helpful for me. This course gave me knowledge about various big data tools. Also helped me to gain knowledge on python and nlp. I got an overview of data analytics. How it works and what are the things we need to master to get more deep into it.