Text

Description automatically generated

**Exploratory Data Analysis (EDA) of Chronic Diseases**

**Project in the U.S.**

**WEEK 2**

**Final Project- Milestone 1**

Logo

Description automatically generated

**Submitted By:** Chailvi Yadav, Liya Ai, Manisha Kaila (Group 1)

**Course Code:** ALY 6010 Probability Theory and Introductory Statistics

**Instructor:** Professor Amin Karimpour

**Submitted on:** November 13rd 2021

**Introduction**

Our group focused on death cases due to chronic disease for a decade (2010-2020) in USA. The purpose of choosing this dataset is to know what is the trend of death cases in USA and where the maximum death cases have happened so that could provide more targeted solutions and treatments for the people who are suffering from the chronic diseases. We have chosen the real-time dataset from data.gov.

**Describe the data fields including the title, the data type, the data description, etc.**

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated with low confidenceFirst, let’s have a look of the overview of the entire dataset. We have approx. 956,638 observations and 34 variables from the datasets. 4 In this analysis, we have prepare and filtered out 5the valid data for analysing due to the missing datasets.

**Describe any data cleaning you did**

We have 60-70k rows of duplicate data which we have cleaned using “ Remove.subset()”. Further, we have also filter out blank columns named ( DataValueFootnoteSymbol, DatavalueFootnote, LowConfidenceLimit, HighConfidenceLimit, etc) using “Remove.subset()”.

**Data Analysis**

*Provide visualizations of the key data and subset data of interest. This should be done for categorical data, discrete data and continuous data.*

🡪 From the below mentioned visualization, the key subset was the Topic of the diseases and how to use that categorical data and converted into frequency count and plot the graphs.🡪

*Provide descriptive statistical tables for key data fields of interest.*

🡪 We have chosen the following fields as it gives us almost overall view of the data.🡪 Graphical user interface, text, application

Description automatically generated

All these data are numerical data. i.e. Yearstart, Datavalue ,LocationId ,Geolocation, and Datavaluealt.

Graphical user interface, text, application

Description automatically generated

Fig: Code to obtain Descriptive Statistics

Table

Description automatically generated

Fig: Output in table form, Created the table using kable:: and knitr()

*Provide analysis above and beyond the graphs and tables. Explain what the tables and visualizations tell you about the data.*

🡪 The data which is taken is a raw data and has a lot of points where in we can check what was the ratio of death cases, even the location and the cause. Topic is a field which tells us the diseases which was the cause of death, year start and End gives us information for how long the diseases was in trend for the particular state.

(1) In the below Scatterplot, we have chosen variables such as Location Id and Start year to get to know the cases in that location in between 2010-2020.

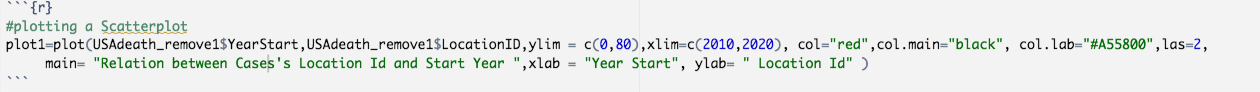


Fig: Code for Scatter plot: Relation between Location ID and Year Start.

Chart, bar chart

Description automatically generated

Fig: Output Scatter Plot.

In this we can see that as 2010 is the start year and the data is till 2020 (initial months) the cases have not been much recorded in these years, and the least cases recorded is in 2018 and the highest in 2017 and 2019.

(2) For the next visualization we did a histogram, first we converted Topic field into a data frame so that we can use the categorical data and create a histogram based on the frequency count and with help of locationID we can see the trend of different diseases.Graphical user interface, text, application, email

Description automatically generated

Fig: Code for creating Data frame and Histogram

Chart, bar chart

Description automatically generated

Fig: Output Histogram

From the output we can see that the locationID near 25-30 have the highest number of cases in this Decade(2010-2020) and lowest is recorded between 55-80.

(3) In the next plot we have created a bar plot, to find the relationship between Location with respective to its Abbreviation to the frequency( count which was created above) to find the cases, the shown plots are for the maximum case cities which were found from all over USA.

Chart

Description automatically generatedFig: Code and output of the horizontal bar plot.

From the above figure we can see that the states mentioned has the highest number of cases of deaths from overall disease topic which is taken into consideration. Out of 50 states 18 states have maximum death cases.

(4) For getting an insight what the cases looked when they ended we used box plot to visualized it.

Graphical user interface

Description automatically generated with medium confidence

Fig: Code for boxplot when the cases ended.Chart, calendar

Description automatically generatedFig: Output of boxplot.

From the above output figure we can understand that the mean is the same for almost all the location except 22 and 38 and 72.

(5) We did one more boxplot to see this time what is the scenario when it is consider start of the year like when the case has started.

Graphical user interface, text, application, email

Description automatically generated

Fig: Fig: Code for boxplot when the cases started.

Chart

Description automatically generated

From the above output figure we can understand that the mean is the same for almost all the location except for Kentucky and Virginia.

**Summary**

In a nutshell, we have taken the raw data from data.gov site, cleaned the data removed null data, then removed the columns which had all null value, changed few data type for fields converted it to numeric so that we can work with the data for visualization. From the visualizations we can understand the mean for the cases in the states are almost same for the whole decade, the cases were recorded less in 2010 and 2020 apart from that 2018 was recorded with less cases.

The outliers in both the box plots are extreme that means the data which is given have lot of noise. Apart from visualization we have also got to learn how to get the data from an open source, how to clean the data and how to work with raw data. From the first glance at the .csv file we can understand this

1. Amongst all the deaths in the United States, 65.8% of deaths are of males and 67.2% deaths are of females as of 2010.
2. Major risk factors include smoking, excessive alcohol consumption, poor diet and insufficient physical activities.
3. From this project we got to learn new skills such as in R about data visualization and how to filter raw data and work on it

To sum up, the data we imported from the Data.gov website is a trusted source of data, it had all data from all over the USA with lots of Raw data including null data as well as unsaturated data. We filtered the data first and then applied conditions in order to get data for a particular region, we selected North-eastern region. The data visualization and the R outputs show us the trend how many such cases are there who dies because of these 15 diseases, it showed that maximum trend was in alcoholism, Cancer and Disability, hence we focused on these 3 topics and continued with our research, From the above visualization figures we can see the maximum death was from prostate cancer in male in the Maine region with data value around 400.

* Amongst all the deaths in the United States, 65.8% of deaths are of males and 67.2% deaths are of females as of 2010.
* Major risk factors include smoking, excessive alcohol consumption, poor diet and insufficient physical activities.
* From this project we got to learn new skills such as in R about data visualization and how to filter raw data and work on it.