Declaration on Plagiarism

This form must be filled in and completed by the student(s) submitting an assignment

Name:	Satyam Ramawat
Student Number:	19210520
Programme:	Masters in Computing (Data Analytics)
Module Code:	CA682
Assignment Title:	Data Visualisation
Submission Date:	13 Dec 2019
Module Coordinator:	Dr Suzanne Little

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. I have read and understood the Assignment Regulations. I have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the source cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.

I have read and understood the referencing guidelines found at http://www.dcu.ie/info/regulations/plagiarism.shtml, https://www4.dcu.ie/students/az/plagiarism and/or recommended in the assignment guidelines

Name: Satyam Ramawat Date: 22nd November 2019

Analysis of Smokers Life Expectancy over the Globe

Abstract (max 200 words)
How much or up to what age smoker male/female adults of all over world lives over their actual life expectancy?

Smoking is practice where substance, dried leaves of alcohol plant is rolled into rice paper and shaped in cylindrical object, which is burned and consumed in form of smoke into our circulatory system. Smoking commonly has negative impact on health of human-being whereas cigarette smoke inhalation leads to challenge physiologic processes of our respiratory system. Injurious to health which derived from smoking leads into shorter life or death.

Conclusion

Smoking is injurious to health, which lead death and reduce life expectancy of human-being. Post analysing data of past years, Male/Female who do smoking having shorter life whereas they die before actual life expectancy.

Overall, Tobacco smoking is popular among many growing countries but it will reduce the life of an individual. So people who smoke cigarette just for fashion or show off or who are addicted should discontinue it because other way around they are killing themselves.

1. Dataset
Data-set has been downloaded from source: https://www.gapminder.org/data/

DATASETS Downloaded		
Datasets	**Location in Folder**	**Actual source**
Life Expectancy Female	Assignment/life_expectancy_female.csv	https://population.un.org/wpp/
Life Expectancy Male	Assignment/life_expectancy_male.csv	https://population.un.org/wpp/
Male Smoker Percentage	Assignment/sh_prv_smok_male.csv	https://apps.who.int/ghodata/
Female Smoker Percentage	Assignment/sh_prv_smok_female.csv	https://apps.who.int/ghodata/

- Each Datasets has 185x151 size.
- Each Datasets has records of 185 Countries and 151 Years.
- Collection of Dataset has belonged to 3 Characteristics of Big Data I.e. *Variety, Velocity and Veracity*. Thus, Different Datasets has been merged and transformed into useful sets.

2. Data Exploration, Processing, Cleaning and/or Integration

- Integrated Life_expectancy_female.csv and Life_expectancy_male.csv and created single Life_Expectancy_Male_Female.csv.
- Integrated sh_prv_smok_female.csv and sh_prv_smok_male.csv and created single Smoker_Male_Female.csv.
- Life_Expectancy Dataset belongs to life expectancy of citizens of country whereas this dataset has all countries life expectancy in Age unit.
- Smoking Dataset belongs to amount Smoking citizens of a particular country whereas this dataset has all countries smoking data in Percentage unit.

- *Life_expectancy_female.csv, Life_expectancy_male.csv, sh_prv_smok_female.csv and sh_prv_smok_male.csv* has records from year 1950 till 2099 so, I've filtered data from the year 2010-2016, merged Life Expectancy Male Female together and Smoking Male Female together by performing calculation.
- Segregated male and female data of life expectancy and smoking has been given in unit out of 100%, when I've merge the value produced out of 200% thus I have added and divide by 2 to narrow down the unit in to 100%.
- ***Life_Expectancy_Male_Female.csv*** and ***Smoker_Male_Female.csv*** are the merged and filtered according to above two points.
- The second level filtered data are available in "*Assignment/Filtered/*",
  Firstly, I've Find list of countries whose Life Expectancy Data is not available in Smokers Data and thus removed those countries and saved as ***"Life_Expectancy_DataFrame_Filtered.csv"***
  Second, I've Find list of countries whose Smoker Data is not available in Life Expectancy and thus removed those countries and saved as ***"Smoker_DataFrame_Filtered.csv"***

- **Smokers_Life_Expectancy_Male_Female.csv** is the final dataset which is created. This dataset has life expectancy of smokers according to country wise.

- **How Smokers Life Expectancy Calculated,**
  From Life_Expectancy_DataFrame_Filtered.csv
  Country = Albania, Year = 2010 has value => 76.8 Age

  From Smoker_DataFrame_Filtered.csv
  Country = Albania, Year = 2010 has value => 31.2 Percentage

  Now, 31.2 is the amount of smoking person's percent from 100 percent whereas 76.8 is the last age to live which is equivalent 100, that mean 76.8 is 100% age in country Albania. So, 76.8 x 31.2 / 100 will give output 23.9616 smoking percentage. Therefore, 76.8 – 23.9616 will give actual age 52.8384 of smokers of Albania. Vice versa, same calculation has been done for all countries and output has been saved in ***Smokers_Life_Expectancy_Male_Female.csv*** and python code is available in "Assignment/ PythonNoteBook /CA-682 Assignment.ipynb" at line [339].

- Used Python Programming Language for data filtration and manipulation, along with libraries and functions:
  - ➔ Pandas
  - ➔ NumPy
  - ➔ Pandas.Dataframe
  - ➔ Set
  - ➔ toList()
  - ➔ drop()
  - ➔ to_csv

3. Visualisation

Tableau has been used for visualization of concepts from data, and file is available at "Assignment/ Tableau / CA682_Satyam_19210520.twb".

1) **Finding out the affected regions who has shorter life from smoking.**
   From the filtered data ***Smokers_Life_Expectancy_Male_Female.csv,*** let's see how many countries has poor and good smoking life expectancy. So in Tableau data has been divided into 2 Clusters whereas used feature from analytics > Model > Cluster.

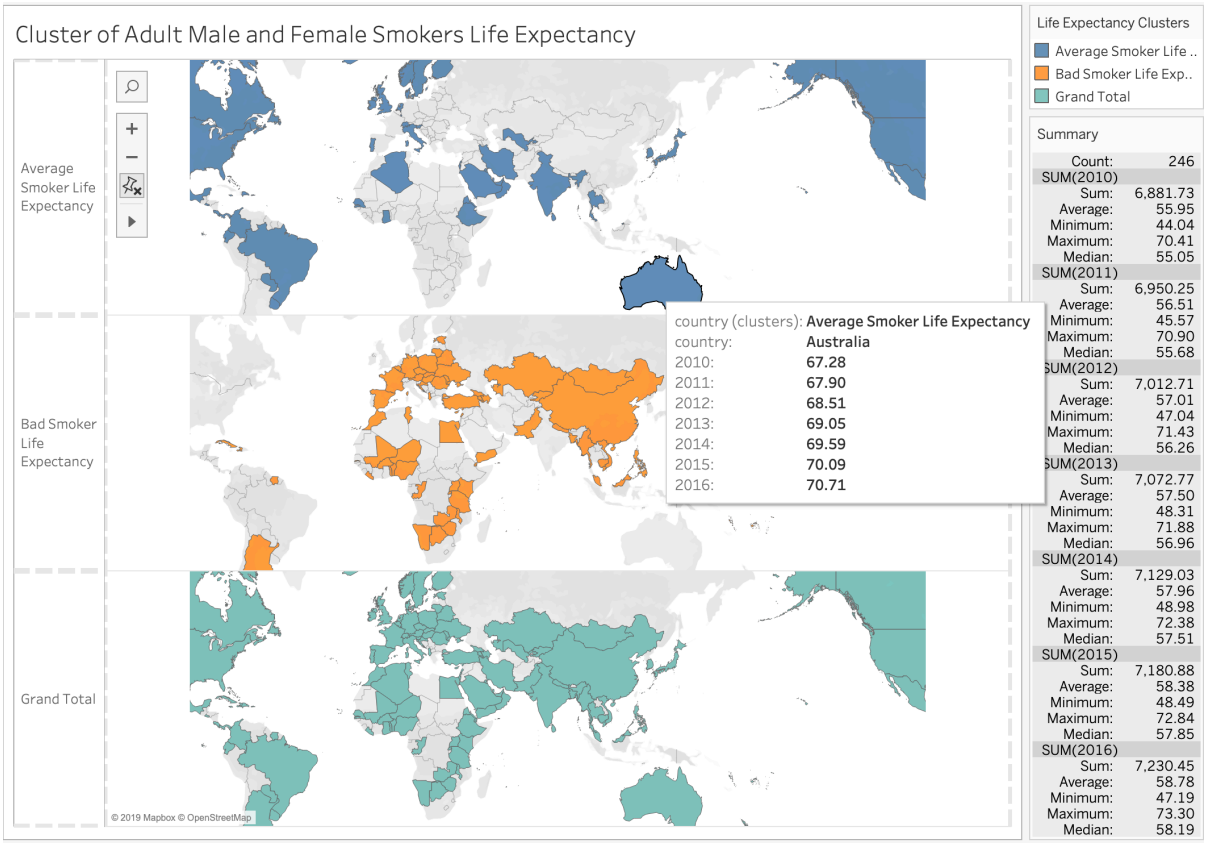
   Cluster 1 defines regions or countries whose smokers has average over good life expectancy, that why they are show in Blue Colour.

   Cluster 2 defines regions or countries whose smokers has poor over good life expectancy, that why they are show in Orange Colour.

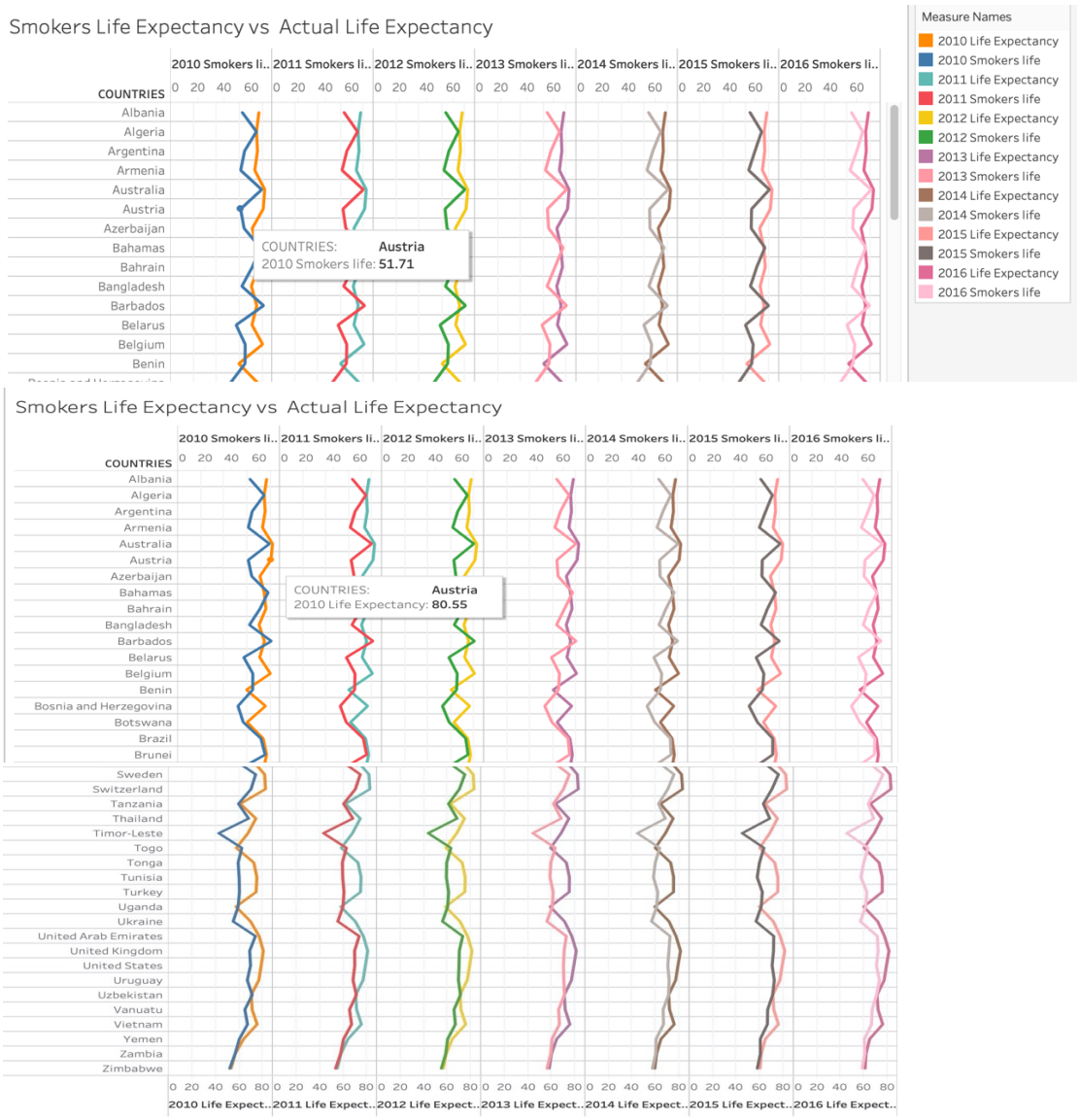
   Overall will signify the affected regions, both average and poor life expectancy from smoking.

   Thus, used **OpenStreetMap Graph** to classify the affected region properly.

   Available at "Assignment/ Tableau /Smokers_Life_Expectancy_By_Region.jpg"



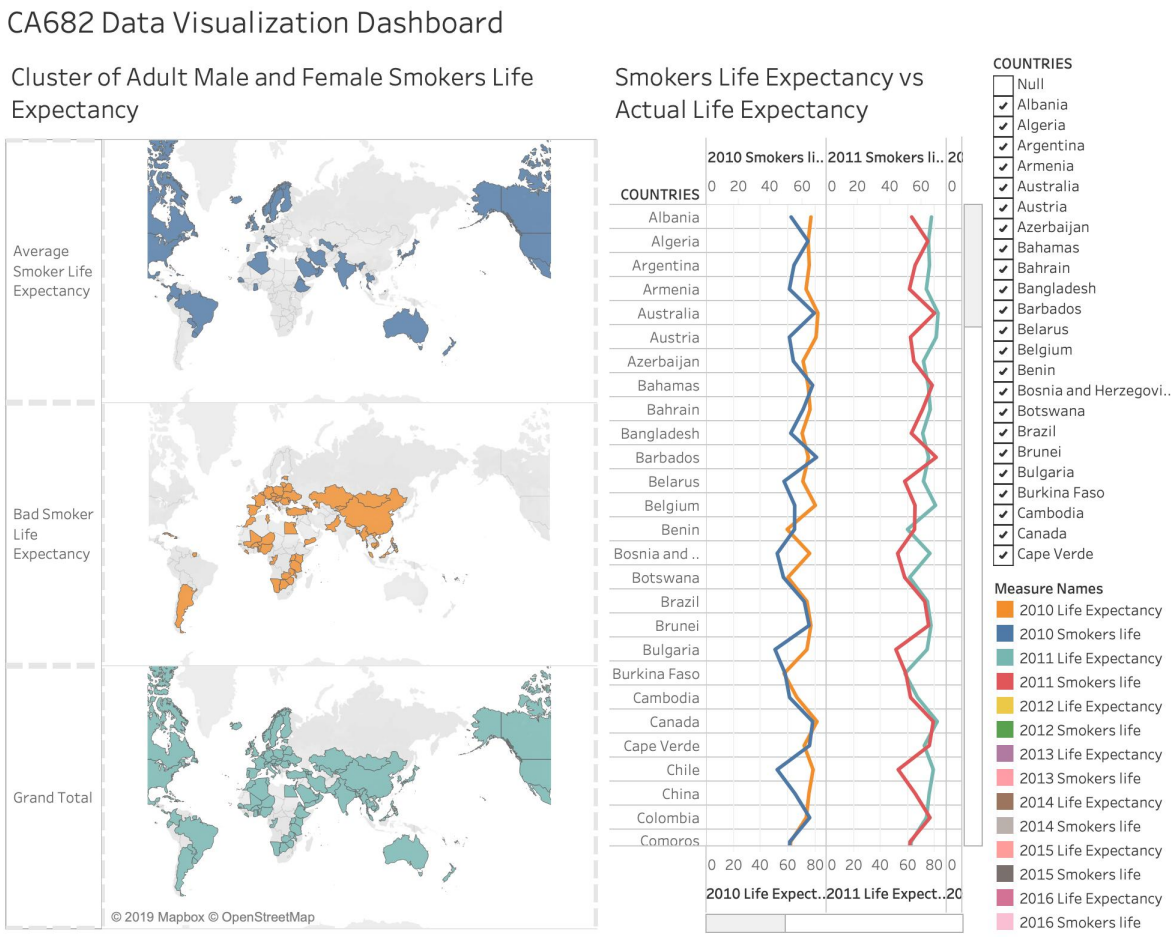
2) **Comparison of *Smokers Life Expectancy* with its *actual Life Expectancy*. To find out how much smoking make a person to live.**
From the Datasets ***Smokers_Life_Expectancy_Male_Female.csv*** and ***Life_Expectancy_DataFrame_Filtered.csv,***



Here, I have used **Horizontal line graph** which is **Dual Axis Graph** for the best comparison, in above graph each year has been shown column wise with countries row wise whereas in a year *Left Line* indicates the life expectancy of smokers of a country and *Right Line* indicates the actual life expectancy which mean without smoking a person can live.

Available at "Assignment/ Tableau / Smokers_Life_Expectancy_vs_Actual_Life Expectancy.jpg"

3) Dashboard – Tableau



****Video is available, and file name is: "CA_682_19210520.mp4"**

4. Technical Challenges Faced

I. Appropriate Data Downloading according to domain.
II. Cleaning and processing four big datasets.
III. Aggregation of Datasets and processing into expected outcome.
IV. Producing of Smoker Life Expectancy from 4 available data.
V. Calculation for age of Smoker Life Expectancy, from smoker male & female percentage and life expectancy approx. age.
VI. Keep data analytics streamed according to domain.
VII. Learning new tool Tableau for Visualization.
VIII. Learning importance of Dual-Axis graph.
IX. Learning about impact of colours in visualizing the graphs.
X. Data Analytics behind Data visualization.

5. **Conclusion**

From the graph 1, we can analyse that countries from Cluster 1 like Australia, India, Ireland and vice versa has good awareness of smoking whereas these countries strictly monitor and restrict the citizens of country to quite or do less smoking, moreover these countries penalized or fine if a person found smoking in a public. Because of strictness people intent to do less smoking as compared to countries of Cluster 2.

From the graph 2, we can see that if a person quit smoking or do not smoke then its life expectancy will rise up. Person who smoke has lesser life, whereas person belongs to good air quality or controlled pollution country but he/she is in a practice of smoking then it doesn't really matter, because smoking damage their lungs directly which more hazardous than pollutant.

Overall, Smoking kills is true phrase, it causes many disease like lung cancer, strokes, COPD, Asthma. It's a slow poison which damage our respiratory system every time we inhale into our body. Damage of lungs / respiratory system leads person to live shorter than actual expected, that's the reason countries from graph 1 intent to save person's life in order to save GDP growth over life expectancy of a country. People or Youngster should involve in Meditation, Yoga or any other sports activity which negates the smoking practice, which results into their better life and future.

References

[1]. https://en.wikipedia.org/wiki/Smoking
[2]. https://www.cdc.gov/tobacco/basic_information/health_effects/index.htm
[3]. https://www.lung.org/our-initiatives/tobacco/reports-resources/sotc/by-the-numbers/10-worst-diseases-smoking-causes.html