

Analyzing and Visualizing Insurance Data

Data Analysis and Data Visualization



UNIVERSITÉ
BISHOP'S
UNIVERSITY

Instructor:

Dr. Rami Yared
Bishop's University

By:

Vishwas Hasija
Student ID: 002251121
Master's in Computer Science
Bishop's University
Sherbrooke, Canada

Abstract- Data Analysis is a process which includes inspecting, cleaning, transforming and modelling data with goal of unearthing the useful information, deriving conclusion which can be help in decision making. Where Data Visualization is a mixer of art and science to display important information/data and relation within the data in graphical format

I. INTRODUCTION

This report is guide for the project for the course work CS503B and CS504B Data Analysis and Data Visualization respectively. Main objective of this report is to provide a complete layout how I have reached to the conclusion of the project. Which all libraries I have used.

II. MACHINE LEARNING

1) Machine Learning:

Machine learning is a field of computer science where computers are made to learn and think like human. Learnings are improved over the time by feeding data and information. Data can be fed by inputting data or interaction of machine with real world. Predictions can be made using this learning. “Machine learning at its most basic is the practice of using algorithms to parse data, learn from it and then prediction about something in the real world”-Nvidia. Fundamental goal of machine learning to generalize beyond training sample and interpret the data that it has never seen. Machine learning is the combination of Representation, Evaluation and Optimization. Where choosing a representation for a learner is standard with choosing a classifier that it can possibly learn. This is called hypothesis set and learner cannot learn without hypothesis set, it is called **Representation**. Each algorithm has implicit evaluating function for itself which is completely different from external one. These evaluating function are necessary part of any machine learning algorithm for optimization called **Evaluation**. The **optimization** is the key for the efficiency of learner. Machine learning can be associated on the basis of learning style (i.e, supervised, unsupervised and semi-supervised) or on the basis similarity in form or function (i.e, classification, regression, decision tree, clustering and deep learning etc).

2) Differentiation on the basis of Learning Style:

Algorithm are differentiated on the basis of their learning style. There is three types of algorithms: Supervised, unsupervised and Semi-Supervised learning algorithms. In **Supervised learning**, algorithms is used to map input data to output data using some function. For example there is input variable (Y) and output variable (X), algorithm is used to learn the mapping function from input to output data.

$$X = f(Y).$$

Data is divided in two parts one part being train and another part being test. Train data is used to learn the mapping function between input and output data. Where test data is used to test the mapping function of an algorithm. Predictive output from algorithm is compared with actual output to check the accuracy of algorithm. In ideal scenario an algorithm will be able to determine the class label correctly for undetermined instances.

Supervised learning algorithms are related to retrieval based AI but they can be also able to generative learning model. **Unsupervised learning**, learns few features from the input data and when new data is entered it used to classify this data on the basis of previous learnings. Unlike supervised learning there is no teacher to make it learn. It is used to classify input data to output data on the basis of similarities. It is left on their own decisive power. Models using unsupervised learning learns the relationships between elements in the data or information and classify the raw data (data without labels) without any help. Chat-bots, self-driving cars, facial recognition programs, expert systems and robots are the some high level of programs which either use supervised or unsupervised learnings. **Semi-Supervised** have the power of both supervised and unsupervised. In semi-supervised type large chunk of data is present out of which some of amount of data have output for its input other don't. Machine learning algorithms have found that when labelled data is used in union with unlabeled data can result into efficient learning accuracy over the unsupervised learnings taking without taking cost and time needed for supervised learning to train the model and acquire label data as unlabeled data is inexpensive to acquire . Semi – Supervised learning can be referred to either transductive or inductive learning. The objective of transductive learning is to infer the correct labels for the unlabeled data. Where aim of inductive learning it to map input X to Output Y. For example of photo archive where some of the photos only are labelled as it is very cumbersome to label all of them.

3) Differentiation on the basis of similarity of function:

Approximation algorithm are the algorithm which are used to map output data Y to input data (X) using some function (f). It is used to develop a model using historical data and predict the output for new data. **Classification predictive** modelling is used to approximate mapping function from input data (X) to output data (Y) where output data is labelled or categorized. It is used to analyse the categorical data can be divided into different classes. Classification learning can be further two types binary and multi-classification. **Binary classification** divide the data into “no/yes”, “0/1”. It answers the questions like whether it will “rain tomorrow or not” or “mail is important or not” or “the picture is of cat or not”. Where **Multi classification** is used to categorized data into more than two different categorize and answer the questions like “whether picture is of tiger or cat or lion” and “email in the mail the box is it important or promotional or spam”. **Regression Modelling** it is used to model an approximate mapping function from input data to output data, where output is a continuous real integer or floating point value such as amount or quantity. A regression can have discrete or real input variable. There can be single input and n-number of inputs also. Regression with multiple input is called multivariate regression problem. A regression problem where input are ordered by time is called time series forecasting problems. Regression algorithm can be estimated by many ways but the most common is Root Mean Square Error also known as RMSE. Classification models is used to predict the discrete labels where Regression models is used to predict the continuous values. Classification models can be estimated by their accuracy whereas Regression models are estimated by Root Mean Square Error. These are the some differences between Classification model and Regression Model. On the other hand Classification model can predict continuous values but these values are in the form of a probability for a class label and Regression model can be used to predict the discrete values but these values are the in the form of integer, this is the overlap between the two models.

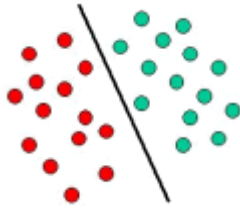
III. DIFFERENT ALGORITHMS

1. SUPERVISED LEARNING

In context of machine learning and artificial intelligence is a type of system in which input and required output are provided. Input and output data both are labelled for categorization to provide the learning basis for future data prediction. It is called supervised learning algorithm because process of an algorithm learning from the labelled data is mimic of a process where teacher supervise learning

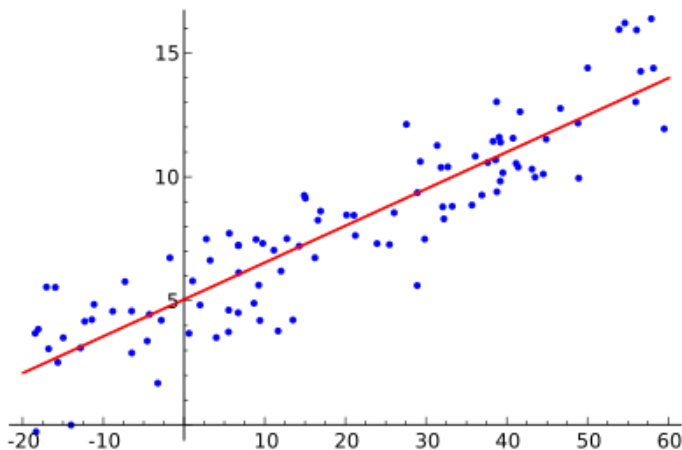
process. When model reaches certain level of accuracy learning is stopped. While considering supervised learning algorithm there are four issues. First issue is **Bias-Variance Tradeoff**, suppose we have multiple good training data set. A model is trained with these data sets, it is systematically incorrect if model predicts correct output for any data set x. Then system is said to be **biased** for x dataset. If model predicts different output values when trained with different input data sets then system is said to have high **variance**. The prediction error of a model is related to sum of bias and variance of the mode. A learning model with low bias should have flexible so that it can fit the data well. But too flexible model, we fit the each training dataset different leading high variance. So Supervised learning model is said to have proper tradeoff between bias and variance. Second issues is **Function Complexity and Amount of Training Data**, amount of training data available relative to the complexity of “true” function. If true function is simple then algorithm with low variance and high bias (“inflexible algorithm”) will be able to learn with small amount of datasets. If the function is complex that is the value of output variable is depend on multiple input data then model will be able to learn with flexible algorithm (high variance and low bias) large chunk of data. Third issue is **Dimensionality of the input datasets**, if datasets have high input feature vector then learning can be difficult even with simple true function for small datasets because it can confuse the model and even cause lead to high variance of the model. If data have high dimension then we need to tune the model for low variance and high bias and this can lead to increase high accuracy of model. Fourth issue is **Noise in Output values**, if desired output values are incorrect often, then learning model should not find the function to match the training data exactly. Attempting to fit the data exactly can lead to overfitting. Overfitting of data can lead to corruption of data sometimes. In several approaches efforts are made to remove the noise in the output values such as early stopping to prevent overfitting as well as identifying and purging the noisy data from the training data sets. Some supervised learning algorithms are :-

Support Vector Machine (SVM): It a biased classifier which formally divide the hyperplane into categories one for each labelled class. This type of algorithm can be used for regression, classifier and outlier detection. SVM is effective and efficient in high dimensional output data. It is still effective even number of datasets is less than the number of dimensions. SVM are versatile as different kernel function can be specified for decision making. These are some advantages for the SVM. On other hand SVM doesn’t provide probability estimates directly, they use five-fold cross validation process which is increase cost and processing time.



Above is the schematic example of linear classification in which it illustrates that the objects belong to either red or green class. Line in between is the boundary for both classes. Right side of boundary we have objects falling under the category green color and on left side we have objects with class having red color.

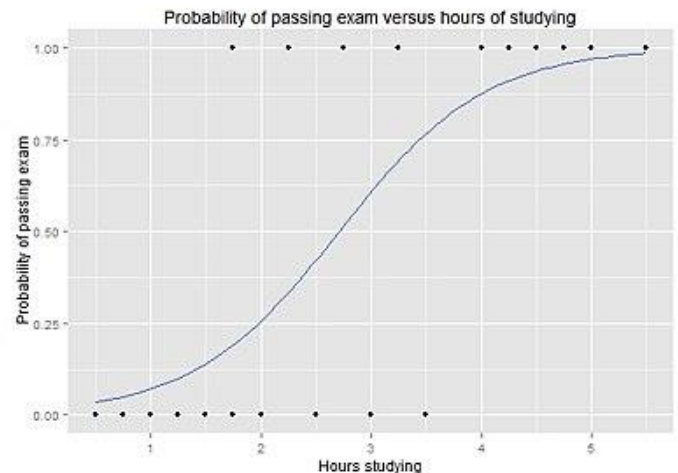
Linear Regression: It is a model to predict the linear relation between one or more explanatory variables (independent variables) and scalar responses (dependent variable). A model with single explanatory variable (independent variable) is called simple linear regression model. Model with multiple explanatory variables is called multivariate linear regression.



Above is the example of simple linear regression model with one explanatory variable and single scalar variable. **General Linear Models** are the models where response variable is not a single variable instead it is vector. They are known as multivariate model but not same as multivariable linear models. **Heteroscedastic Models** are the models which allow heteroscedastic which is error for different response variable have different variance. Weighted least square is a method used for estimating linear regression models when the response variable may have different error variance. **Hierarchical Linear Models** are used to organize data into hierarchy of regressions like A is regressed on B and B is regressed on C and so on. It is often used where variable have natural hierarchy structure like patient fall under

doctor, doctor fall under hospital and hospital fall under a wing or block. Linear Regression model is used in biological, behavioral, and social science to unearth the possible relation between the corresponding elements.

Logistics Learning Model: Logistic model is widely used in statistical modelling to model a function to binary scalar response variables. Mathematically binary logistic model has scalar response variable with two possible outcomes like Win/lose, True/False, Pass/Fail, Dead/Alive. These values can be scaled down to 0/1.

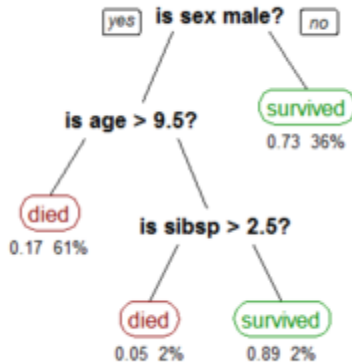


Above graph depicts the regression curve showing the probability of passing in an exam on hours of studying.

Naïve Bayes Classifier: It is a family of simple probabilistic classifier based on Naïve Bayes theorem with naïve independent assumptions between the features. Naïve Bayes Classifier are highly climbable, requiring number of parameters linear in the number of variables in a learning model. Naïve Bayes is a technique for constructing classifier models that assign class labels to problems, represented as a vector of features, where labels are drawn from finite set of data available. There are multiple parameter estimation models related to Naïve Bayes Classifier. First, **Gaussian Naïve Bayes** when we are dealing with continuous values a typical assumption is that the continuous values associated with each class are distributed according to Gaussian distribution. Second **Multinomial Naïve Bayes**, samples are in the form of frequencies with certain event has generated by multinomial (a_1, \dots, a_n) where a_i is the probability of event i .

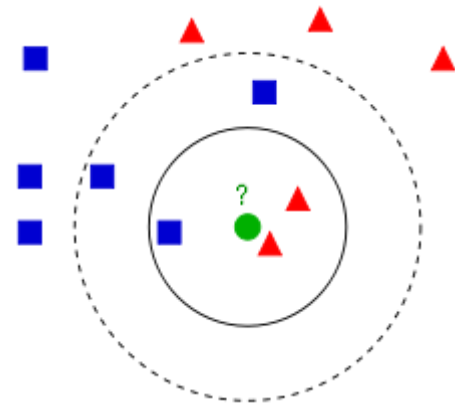
Decision Tree Learning: Uses predictive model to go from independent variable to a scalar response variable about the target value. Tree model where scalar response variable can take discrete values are called classification tree where scalar response variable take continuous values are called

regression tree. Leaves are represented as labels or continuous values and branches are connection between independent variable and dependent variable describing relation between two.



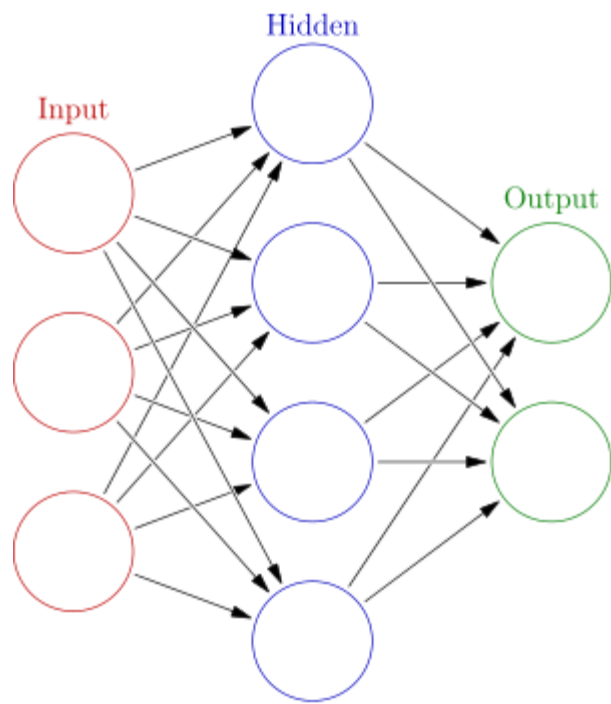
A Tree showing the survival chances of passenger on titanic. Advantages of decisions trees are they are simple to understand and interpret. As trees can be represented in graphical view it makes easy for some to understand the decision tree. They are able to handle both numerical data and categorical data. Other techniques are used to analyze one type of variable. They need little data preparation for analyze while other models required data to be normalized first then used. These models perform well with large set of data using standard computer resource with efficient time in processing. It mirrors human decision making capabilities. Limitation of decision tree are that they can't be robust that is small change in training data can lead to large change in tree and consequently affects the final prediction. Decision tree models can create large complex tree which doesn't go well with from the training data.

K-Nearest Neighbors algorithm: Also known as K-NN algorithm. It is a non-parametric model used for classification and regression problems. In K-NN classification dependent variable is a member of a class. An output is decided on the base of popularity vote of its neighbors with the object being assigned to the class most common among its K nearest neighbors. In K-NN classification dependent variable is the property value for the object. This value is average value of K' nearest neighbors.



Above pictorial represent shows K-NN classification. The test sample (green ball). It can be either classified to the first class of square in blue or second class of triangle in red. If $K = 3$ it is assigned to class 2 in red. If $k=5$ it is assigned to class 1 in blue.

Artificial Neural Networks: Also known as connectionist system inspired by the biological functioning of human brain. Neural network is a combinational framework for different machine learning algorithms which work collectively to process complex data. Such systems tend to learn to perform estimation by considering examples without being programmed with any specific rules. Example in image recognition model might learn to identify images that contain rat by analyzing example image of rat which were labelled manually.



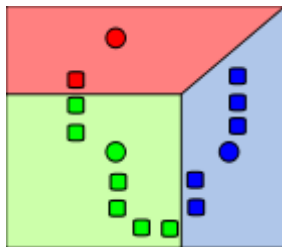
A typical example of neural networks where group of nodes are interconnected similar to the vast network of neurons in

human brain. Here each circular node represent artificial neuron and edges between them is connection from one neuron (output) to other neuron as input.

2. Unsupervised learning

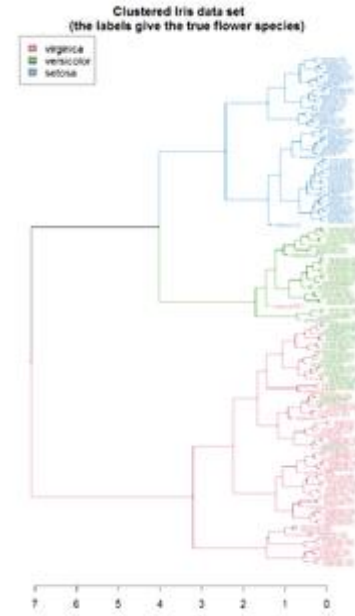
Unsupervised learning algorithm are part of machine learning algorithms and approach which works “no-ground-truth” data. Best way to understand machine learning algorithm is, consider you give exam and you are marked on the basis of the closeness of your answer to the actual answer. What if there are no actual answers to grade you. Then you would be marked on the basis of estimation/approximation from independent to dependent variables. It can also be seen as learning without teacher also known as self-organization. Below are the type of unsupervised learning. Finding right algorithm and hardware is one of the major problem in unsupervised learning. Wrong algorithm and hardware can loss into form of time. It is also difficult to figure out that job is really getting done or not. As there is no label in unsupervised learning algorithms, it almost impossible to measure the accuracy of the algorithm. Last but not least it depends upon data whether an unsupervised algorithm will work on or not.

Clustering Algorithm: Clustering algorithm process the data and try to find out the natural clusters that occurs within the data. One can modify the number of cluster in which they want to divide data in which let one to decide the granularity of the groups. Within Cluster Algorithm we have three types of clustering algorithm K-means clustering, Hierarchical clustering and probabilistic clustering. **K-means clustering**, initially data sets are clustered into the number of data sets. Then K means try to cluster n data sets into k clusters with each observation belongs to cluster with nearest mean. This algorithm is NP hard.



3 clusters are formed by associating each data sets with nearest mean. Each division is represented by Voronoi diagram. **Hierarchical clustering** is a method cluster analysis which used to cluster data sets into parent-child clusters. There is two type of hierarchical clustering. First, **Agglomerative** it is a bottom-up approach where each data point is clusters into each cluster as we move up the hierarchy we pairs of clusters are merged up. It can be also viewed as the concept of generalization. Second is **Divisive**,

it is top-down approach were data sets are clustered together and as we move down the hierarchy clusters are split down. This process can be seen as specification.



Hierarchy clustering dendrogram of iris datasets.

Third and last is **Probabilistic clustering**, data points are clustered into clustered based on the probability of occurring.

3. Semi Supervised Algorithm: Semi supervised learning models are the class of machine learning algorithms where model is trained by using small amount of labelled and large chunk of unlabeled data. That is why they are called semi supervised algorithm as they are mixture of both supervised and unsupervised learning algorithms. Researchers have found the when labelled data is used with unlabeled data then model have produce great improvement in terms of accuracy as compare to supervised and unsupervised models alone. As unlabeled data is inexpensive in acquisition and it require manual effort or experiments for acquisition of data. Semi supervised models seems to be of great use and practical value.

IV. DATA ANALYSIS

Data Analysis is initial component for data mining and business intelligence. It is a key component to gain insight of data and help in concluding decisions. Various organization these days analyze data from tools like Big Data management solutions and customers experience management solutions that transform data into actionable insights. Data analysis is the field which seek all three past, present and future to answer questions like what happened, what is happening and what will happen. It is transforming data, question and answer phase which leads to important decision making phase in the overall phase of business intelligence. Seven key steps that are involved in process of data analysis in an organization:

- **Decide on the Objectives:** Objectives for data analysis team are determine to develop a qualitative and quantifiable way to decide whether business is progressing towards its goals or not.
- **Identifying business levers:** Identify business levers, metrics and goals in the project to focus on scope of data analysis. It means an organization should be willing to changes key process and metrics to reach it goal.
- **Data collection:** Data is gather from as many as sources possible to build an efficient model and gain more actionable insights.
- **Data cleaning:** Quality of data is improved to get correct results. Incorrect conclusions are avoided. Try to automate this process but employees are involved in to oversee the data and remove inconsistency.
- **Grow data science team:** Increase size of team with individuals who have proficiency in statistics who will focus on data modelling and analysis. Equip them with large scale analysis platforms required.
- **Optimize and Repeat:** Keep refining process and model and repeat process to generate accurate results.

Consideration and Issues in Data Analysis:

- **Having necessary skill to analyze:** Investigator can defer the over the selection of method used by research team analyst. In ideal situation investigator should have more than basic understanding of the process for selecting one over the other.

- **Concurrently selecting data collection methods and appropriate analysis:** Optimal stage for determining appropriate analytical methods occurs during the early stage of data collection and should be consider thereafter.
- **Drawing unbiased inference:** Main aim of analysis is to differentiate between an event occurring as either true effect versus false one. Any bias occurring in training data or selection of model of analysis will lead into increase in chances of drawing biased inferences.
- **Lack of clearly defined objective and outcome measurements:** Poorly defined objective outcome of measurements can lead to increase in likelihood of murky interpretation of the results. This can be done either intentionally or by poor design.
- **Provide honest and accurate data analysis:** The basis for this issue is to reduce the statistical error. Common challenges include removing outliers, fill in missing data and altering changing data.
- **Training of Staff conducting analyses:** Data integrity is at risk when analyst have received inconsistent training. One analyst can assign different rating to data as compared to other. There used be proper training, routines checkups and proper protocol used be maintained with proper analyses techniques.
- **Data recording method:** Analyses could be impacted by the method of recording data. For example:
 - Recording audio or video and transcribing them later
 - Open ended or closed ended survey
 - Prepared field notes by observer/participants
 - Self-administered surveys

While each method have its advantages and disadvantages. Objectivity and subjectivity issues can be raised when data is analyzed.

Data Analysis is broken into four types:

- **Descriptive Analytics:** Using numbers and figures it describe how what has happened over a given

period? How the numbers have changes have they gone up or down

- **Diagnostics Analytics:** It is used to find the reason why something has happened. It involves more scattered data to answer questions.
- **Predictive Analytics:** As the name suggested it moves to foresee what's going to happen in future. How sales will be in future
- **Prescriptive Analytics:** It suggest course of action in future on the basis of some condition.

Exploratory Data Analysis: Also known as EDA. It is an approach to analyze data. It is a technique where analyst try to get first view of data and tries to make sense of it. It's is often first nail towards analysis, used before any statistical techniques are applied. EDA is not a set of process or paradigm, rather it is termed as "Philosophy" in Engineering Statistics Handbook. It helps analyst to get touch and feel of data. It helps to determine the essential features in the data.

Purpose of EDA:

- To check for mistakes and missing data.
- Gain as much as insights of data and underlying its structure.
- Confirms assumptions associated with any model fitting the data.
- Enlist anomalies and outliers.
- Identify the features which have great influence.

Data Analytics tools can be of great use they can bring data to real existence (being of some importance) and they can deliver some value to customer. Lot of hard work is required to extract and transform data but once it is done it can be of great use and it can bring broader insights to customer, business and industry.

There are three broad categories of analytics that offer insights at different levels:

- **Traditional Business Intelligence:** Provide traditional and recurring reports.
- **Self Service Analytics:** Enable users to formalize their own analyses with data and tool provide.

- **Embedded Analytics:** Provide business intelligence of traditional business application like HR systems, Enterprise Resource Planning or Customer Relationship Management. These analytics provide decision making support to user.

There are various tools present which can make our work easier and efficient.

Tableau: With vigorous functionality and high speed to insight. Tableau can be connected to different local and cloud based data sources. Tableau interface provide data sourcing, data preparation, analysis and presentation in a streamline workflow. Tableau's versatility makes it suitable for all the above categories. Tableau robust server house the recurring reports easily. Tableau can be easily integrated with JavaScript API and single sign-on functionality.

Looker: Looker emphasize on providing unified data environment and centralized data governance with more weightage given on reusable components for data users. With extract/load/transform (ELT) approach, Looker give user functionality to model and transform data as they need it.

Looker reusability component provide components for data connections, analysis, visualization and distribution. Looker can be easily integrated with tools like Jira, Slack and Segment.

Dataiku: It combines data analysis lifecycle into one tool. It enable analysts to source and prepare data, integrate with data mining tools, develop visualizations for end users and set up on going data flows to keep visualization fresh. With its focus on data science, Dataiku tends to serve deeply analytical uses cases like churn analytics, fraud detection and demand forecasting

RapidMiner: RapidMiner emphasizes on insights of complex data efficiently. Its graphical user interface includes pre-built data connectivity, machine learning components and workflow. It can be integrated with Python and R, RapidMiner automates all the process involved in data analysis. This platform also speed up the back-end process with environment collaboration and integration with Hadoop and Spark big data platforms.

Domo: Domo focuses on speedy insight for less technical expert people. Its vigorous intelligence

capabilities enable visualization commenting to facilitate collaboration. Domo also provide native mobile support with same functionality as desktop.

Birst: Birst focuses on solving one of the most trivial problems in data analytics. Birst user's data tier automatically provides unified view on data to users by sourcing, mapping, and integrating data sources.

Areas where Data Analytics Application have been employed:

- **Policing/Security:** Several cities have used predictive analysis in predicting areas that would see increase in crime like Chicago, London, Los Angeles. Though it doesn't make possible to catch suspect. But, yes it has help in reducing crime rate by deploying police officers within certain areas.
- **Transportation:** During London Olympics, TFL and train operators ensured smooth journeys by using data analytics. They were able to input data from events that took place and forecast a number of people that were going to travel.
- **Fraud and Risk Detection:** This being one of the initial applications of the data analysis. Many organizations have bad experience with bad debt and false claims. Since financial institute are gathering data from very beginning it help them to mitigate the loss using data analysis on the information present for customers each transactions.
- **Manage Risk:** For any organization providing insurance foremost task is risk management. Using data analytics insurance industry gather all the information like claims data, actuarial data and risk data covering which help in making decision. It helps underwriter to perform evaluation and set appropriate insurance charge. These days analytics software is used for detecting fraud claims and raise claims.
- **Delivery logistics:** Well analytics have no limited application. Major courier industries like FedEx, UPS, DHL uses data analysis to make their services efficient and effective. Optimal delivery routes, best delivery data are found using data analysis.



- **Web Provision:** In today's fast pace world, smart cities are equipped with high speed internet provided by either companies or government. Providing high speed is one thing, providing right content and providing internet at right place and right time is another. Like in weekdays organizations and business center need high bandwidth while in weekend it should be shifted to residential complexes. This transition seems to be simple but it is trivial. But it can be achieved by data science and analysis.
- **Proper Spending:** Another problem is that Smart cities large chunk of money on small amount on small work. This could lead to scarping of some important projects. Data Analytics can help to target the projects and help in using tax payer's money wisely. The target of spending money wisely would lead to facelift all the infrastructure using optimal amount.

- **City Planning:** One of the biggest mistake is not considering analytics for city planning. Web traffic and marketing are still being used instead of creation of spaces and building. This causes real issues while zoning and amenity creation. Models that are built will maximize the accessibility of specific areas and without overloading the basic amenities.
- **Travel:** Data Analysis help in optimization of traveler's experience via social media and mobile/weblog data analysis. This is because customers preference and desires can be obtain from this. It will help companies sell product from correlation of the current sales to recent browse history and gain profit.
- **Internet/Web Search:** There are many search engines like Google, Bing, Yahoo, AOL and etc. Each of these search engine are based on Data science application because they use algorithms to deliver the best results for any search directed the results in just split seconds.

V. Data Visualization

Data Visualization is the pictorial representation of data and information present which can help in shaping decision and analyze the current scenario of an organization or industry. By using visual components like graphs, charts and maps using data visualization tools that help to understand trends, outliers and patterns in data.

Our eyes are drawn towards colors and pattern. Eyes distinguish between blue from red, circle from square. Data visualization is an art with perfect blend of technology with it. It is easy for an eye to catch outlier while easing data or information in shapes or colors. One can easily relate to this if he/she has seen big chunk of data in spread sheets. It is true the graphical representations are easy to interpret.

Some general type of data visualization components:

- **Charts:** It represent the information in the tabular form. Some of the examples are menu of a restaurant, rules at swimming pool.

- **Tables:** It lets us visualize data in metrics format. It is also known as data table or data grid.
- **Graphs:** It represent the information as a series of coordinate displayed on multi-dimensional axis. Each value at co-ordinate is related to others through some mathematical relation.
- **Maps:** Maps display geographic information in the form of shape, path, and symbol elements.
- **Infographics:** Infographics are visual representation of facts, events or numbers that reflect patterns and align to story.
- **Dashboards:** Is a collection of resources assembled to create a single unified visual display.

Specific methods to visualize data:

- Area Chart
- Bar Chart
- Bubble Cloud
- Bullet Graph
- Cartogram
- Circle View
- Dot Distribution Map
- Gantt Chart
- Heat Map
- Histogram
- Matrix
- Network
- Polar Area
- Radial Tree
- Scatter Plot

- Text Tables
- Timeline
- Treemap
- Wedge Stack Graph
- Word Cloud

Tools for data visualization:

- **Tableau:** Is a big data visualization tool. Tableau enables you to create charts, graphs, maps and etc. It is available as desktop application and it also provide cloud service.
- **Infogram:** It enables you link their visualization and infographics to real time data. In 3 simple steps many templates can be personalized with their visualization like charts, graphs, maps and etc.
- **ChartBlocks:** It is an online tool easy to use. No coding is required while creating visualization. Graphs can be drawn while using spreadsheets, databases and live feed too. Graphs will be responsive and compatible on all screen size and will be compatible across platform.
- **Plotly:** It helps in creating sharp and slick chart in just a few minutes using spread sheet. Organizations like Google, The U.S Airforce and The New York University are using Plotly. It is user-friendly web tool. It provide integration with JavaScript and Python.
- **Visual.ly:** Visual.ly provides services in visual content. They have big data visualization services. Nike, Visa, Twitter, Ford and the National Geographic are few of the organizations which uses services of Visual.ly. Visual.ly can be trusted if you want to outsource your work to a third-party where you can describe your project with them and can be connected with their creative team.
- **D3.js:** D3 stands for Data Driven Documents. It is one of the best

visualization library that provides integration with JavaScript and uses HTML, CSS and SVG.

- **Chartist.js:** It empower Sass and provide full customization. It can be integrated with Angular JS, React, Meteor, Ember and WordPress.

Seven Benefits of using Data Visualizations:

- **Faster Action:** Our Human brain interpret visuals faster and easily than written information. Use of graphs and charts to summarize data ensures faster interpretation of relation than clustered reports or spreadsheets. This act as clear form of communication between the business leader, stake holders and help them in shaping the future policies.
- **Communicate findings in constructive ways:** Sometimes business reports submitted to higher management are formalized documents filled with graphs and table. Reports become so monotonous that it don't make impact on those whose decision matters. Data visualization tools can make it possible to capture important details and make it easy to interpret these details.
- **Understand connections between Operations and Results:** Visualization facilitate user to track connections between operations and business performance. It is very essential to see relation between market performance and business functions.
- **Embrace emerging trends:** With large chunk of data collected from users can expose to many opportunities for adaptable companies. Collecting information and analysis should be iterative process. By using visualization key factors can be taken into consideration it can help leaders to spot market shift and trends easily.
- **Interact with data:** With data visualization changes can be exposed in timely manner. Interactive data visualization encourage users to explore and manipulate data to uncover relations and factors.

- Create new decisions: With data visualization one have all means ready to tell the stories from the data. Visualization enables to tell the development of product performance in different region and can help down to track what went wrong or well. It also allows leaders to drill down into reason of performance.
- Machine Learning: Come One, Come All: It is not that only large organizations like Google, Apple, Yahoo, Amazon are using machine learning, today almost all organizations are using machine learning.

VI. IMPLEMENTATION

Data Analysis

Objective: Main objective of this project was to analyze data for an Insurance Company XYZ financial. XYZ financial wants to get an algorithm modeled for them to estimate the Cost of Insurance (COI). XYZ decides the cost of insurance based on certain factors like Age, Gender, Body Mass Index(BMI), Number of Children, Smoker Status and Region. Alongside a model they also want a desktop application with Graphical User Interface with capability of taking input from users and estimate the cost of insurance.

Software Requirements:

- Language: Python3.7
- Integrated Development Environment: Synder3
- Libraries: Tkinter, Numpy, Pandas, Matplotlib, Sklearn

Machine learning Algorithm:

- Supervised Regression Algorithm
 - Linear Regression
 - Polynomial Regression

Implementation:

- Data Sourcing: One of the foremost step of this project is acquisition of data. Kaggle one of the sites which is known for datasets

for machine learning algorithm. I acquired data from kaggle.

- Once data was obtained, it was the time to decide which algorithm will fit the data best and give the best result.
- But before that data needs to be in form of machine readable format as machine already understand numbers.
- In data for some features data was not in numeric format fields like Gender, Smoker Status and Region. Data was scaled to numeric format so that it can be in machine readable format.
- After that data was divided into ratio of 80% and 20%. 80% data was used to train the model. While 20% was used to test the trained model.
- Once data was scaled down, firstly Linear Regression Algorithm was used to model algorithm. While using Linear Regression Algorithm maximum accuracy was 69%. Which was not up to the mark.
- Later Polynomial regression algorithm was used to train the model. With Polynomial regression algorithm accuracy was 89% keeping dependent variables 5. Accuracy can be increased further by increasing the number of dependent variables (this is not always true).
- Then I used Tkinter library for developing Graphical User Interface (GUI) as per the requirement. Tkinter is a python framework which enables one to develop desktop application.
- With Tkinter I enabled my application to take input from user and estimate the cost of insurance using the modeled algorithm.
- Desktop application also display's some facts about the data.

Why linear Regression Algorithm?

Linear Regression is a machine learning algorithm based on supervised learning. It performs regression task. Being the primary algorithm in supervised regression learning, first I decided to train the model using linear regression. Linear Regression predicts the dependent variable (Y) based on given independent variable (X).

$$Y = a + b.X$$

Where:

Y is output variable/labels

X is univariate input variable (one input variable)

a is intercept

b is coefficient of X

Relation between a and b and input variables is known as regression model. Primary objective is to find estimated values a and b which would provide best fit for data points. While using

Disadvantages of Linear Regression

- Linear Regression algorithm assumes that there is straight line relationship between the features.
- It is very sensitive to anomalies and one outlier can impact the learning of the model.
- Also if number of sample data is more than the number of features than model starts modelling noise.

Few of disadvantages have led to 69 % of accuracy of the model. Which is not upto the mark. Then I decided to train the model using Polynomial Regression Model.

Why Polynomial Regression Model?

In Polynomial Regression relationship between the dependent variable(Y) and independent variable(X) is modelled as a degree of nth polynomial. Polynomial Regression fits a nonlinear relationship between the dependent (Y) and independent variable (X).

$$Y = a + bX + C.pow(X,2)+e$$

As the relation between independent and dependent variable is polynomial model has given good results. Accuracy of 89%. For polynomial regression analysis data is first scaled to numeric format. Then after scaling data, it is separated into two parts one is used for training and one for testing purpose. Original features are converted to high order using PolynomialFeatures class from scikit-learn. Degree of polynomial feature was set to 5. More the degree of polynomial feature more the accuracy is.

Data Visualization:

Objective: To represent information in form of graphs, charts and maps. And To facilitate leaders to understand the relation between different features and shape their decisions.

Software Requirements:

Language: Python3, HTML, CSS

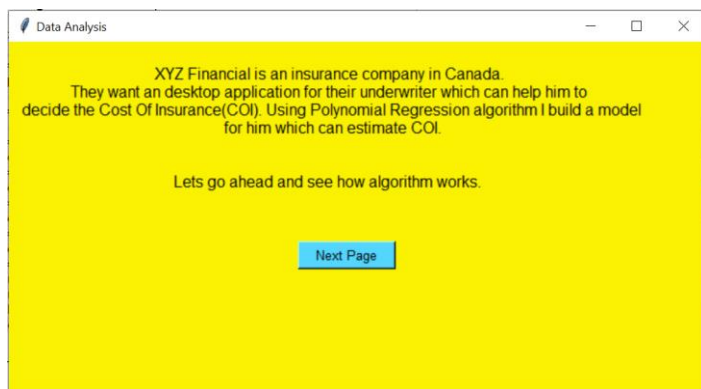
Integrated Development Environment: Spyder3

Libraries: Matplotlib, Seaborn.

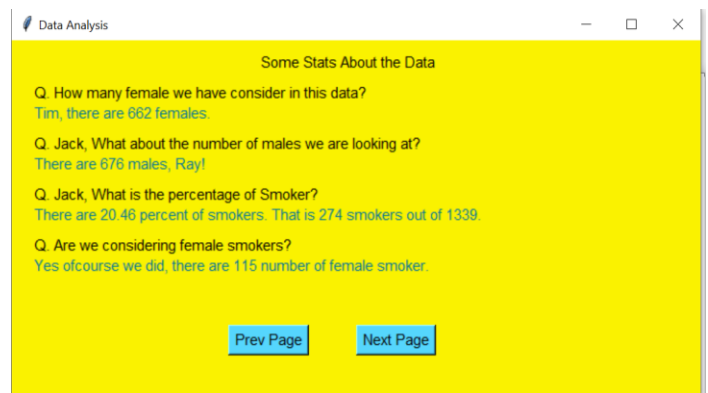
Implementation:

- I used Matplotlib and Seaborn to represent data and information into graphical view.
- Information is displayed using various graphical components like Pie chart, Bar graph, Grouped Bar graph, Scatter Plot, Line Graph, Donut Graph and Heat Map.
- Each components have its own significance.
- All the graphs was displayed on web pages which are interlinked to each other. Each web page contain one graph with their introduction.
- Web pages were designed by using HTML and CSS.

Screen Shots



Welcome Page of Desktop Application for Data Analysis



Stats about the data.

Age	
Gender (M/F)	<input type="radio"/> Female <input type="radio"/> Male
BMI	
Number of children	
Smoker (Y/N)	<input type="radio"/> Smoker <input type="radio"/> Non Smoker
Region	<input type="radio"/> NE <input type="radio"/> NW <input type="radio"/> SE <input type="radio"/> SW
Accuracy	
Estimated COI	

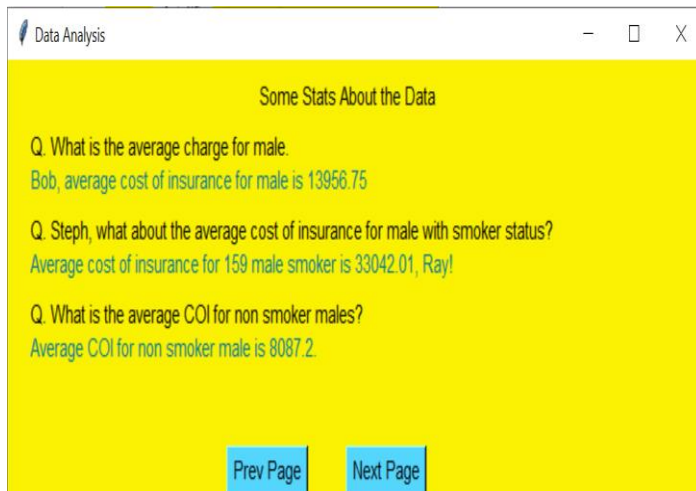
[Submit](#) [Rest Values](#) [Next Page](#)

Page where Underwriter can enter input and see the cost of Insurance

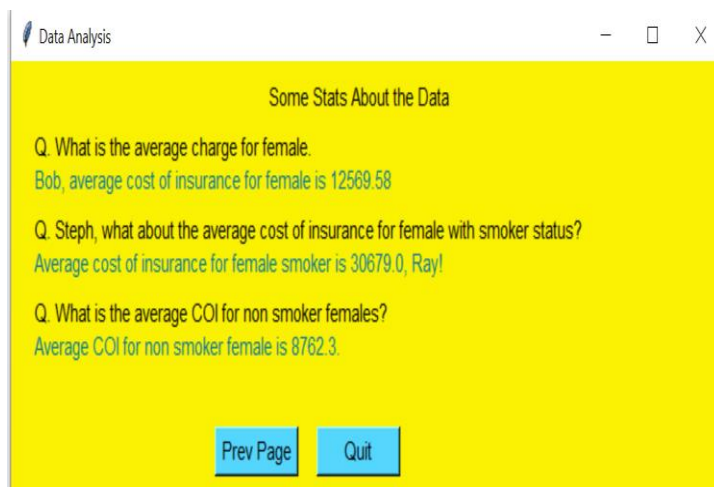
Age	23
Gender (M/F)	<input checked="" type="radio"/> Female <input type="radio"/> Male
BMI	20
Number of children	1
Smoker (Y/N)	<input checked="" type="radio"/> Smoker <input type="radio"/> Non Smoker
Region	<input checked="" type="radio"/> NE <input type="radio"/> NW <input type="radio"/> SE <input type="radio"/> SW
Accuracy	89.32%
Estimated COI	\$14074.0

[Submit](#) [Rest Values](#) [Next Page](#)

One Underwriter enter the values, he clicks “submit” button. Accuracy and estimated COI are calculated upon the calculation. “Rest Values” helps to clear all the value and ready for new input.

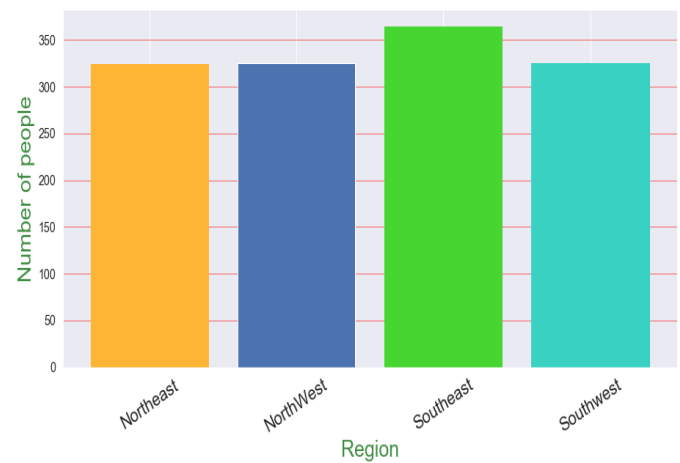


Stats about the data



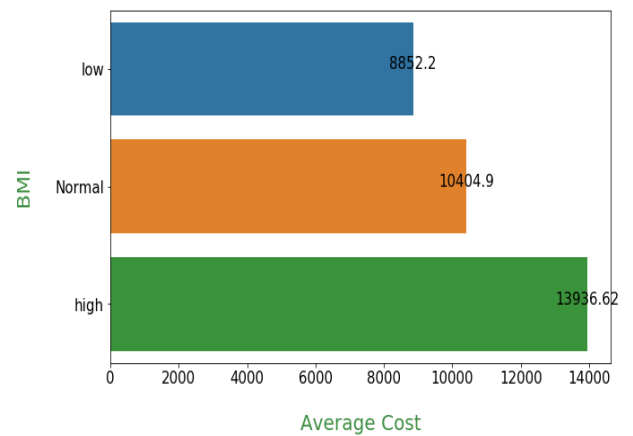
Stats about the data

Region wise Policy Holders



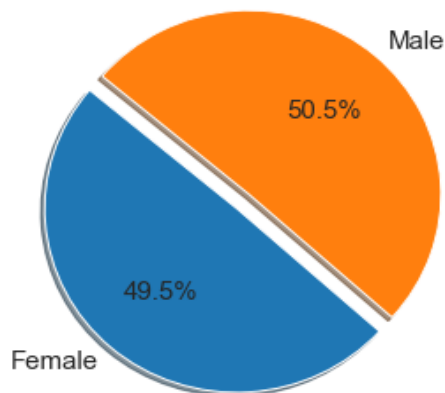
Region wise Policy Holders

Average charge per BMI group



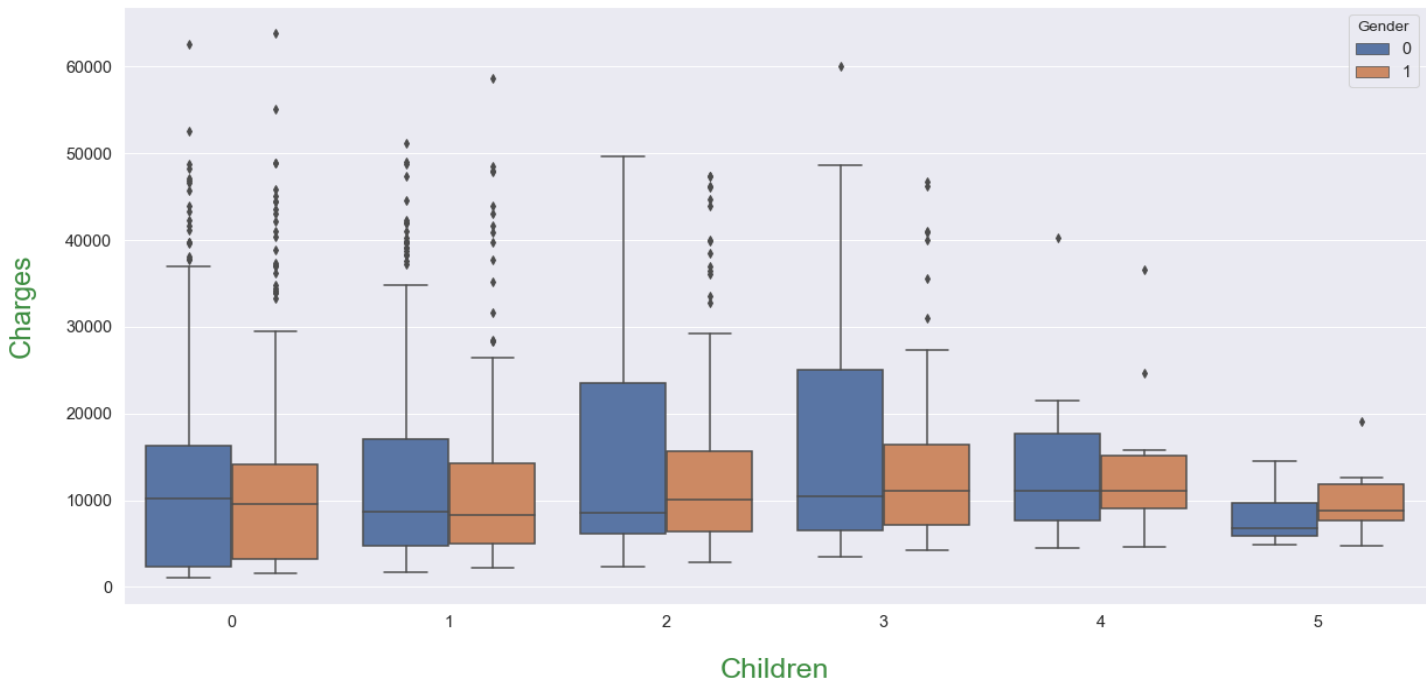
Average Cost Per BMI Group

Percentage of Male and Female

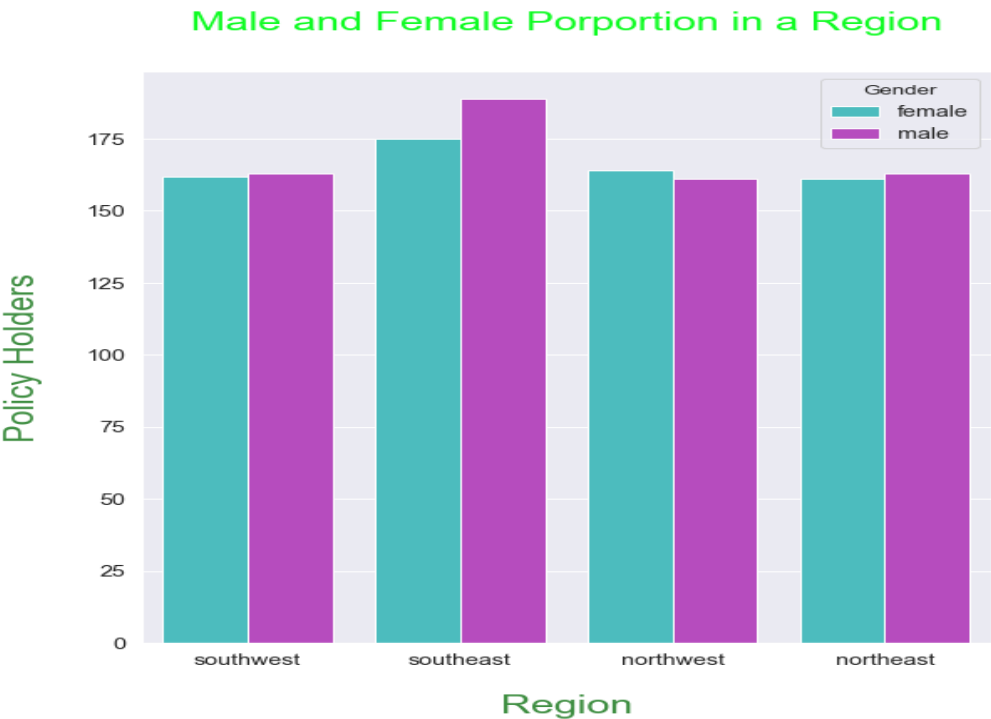


Represent Percentage of Male and Female.

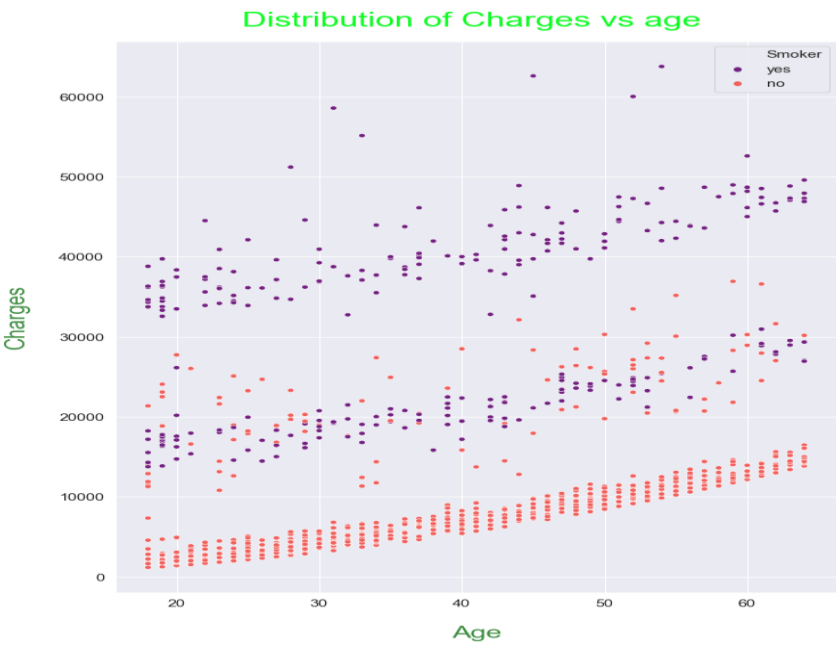
Relation between Charges and number of children



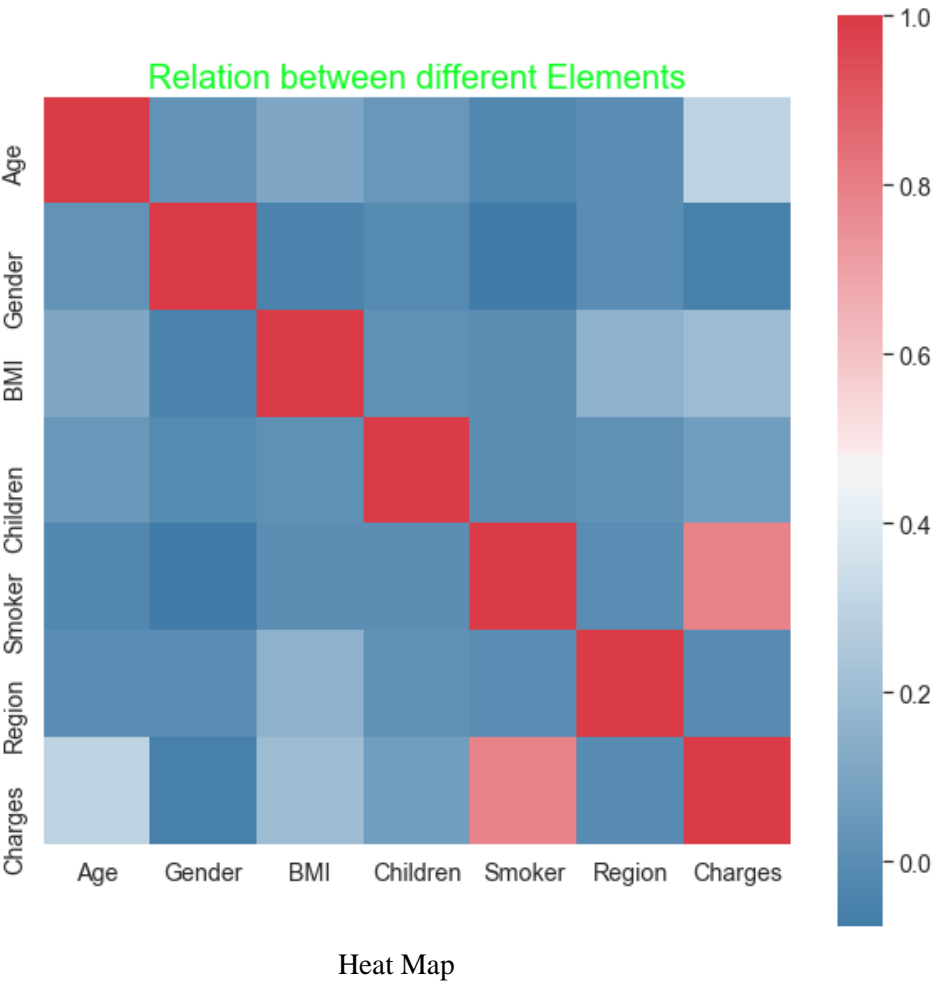
Cost of Insurance as per the number of children for male and female separately



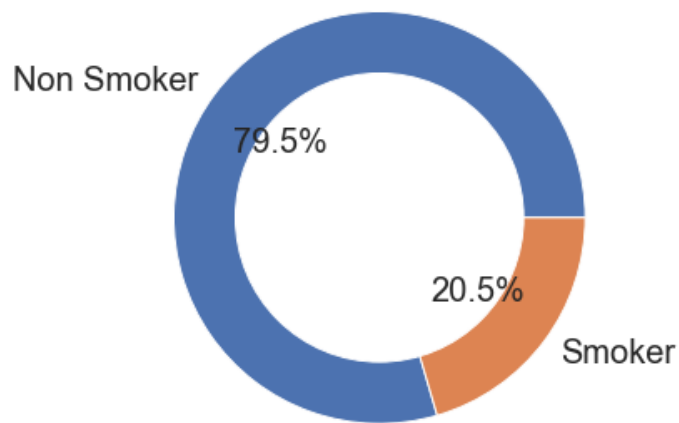
Male and female policy holders in each region



Cost of Insurance according to Age of Smoker and Non Smoker

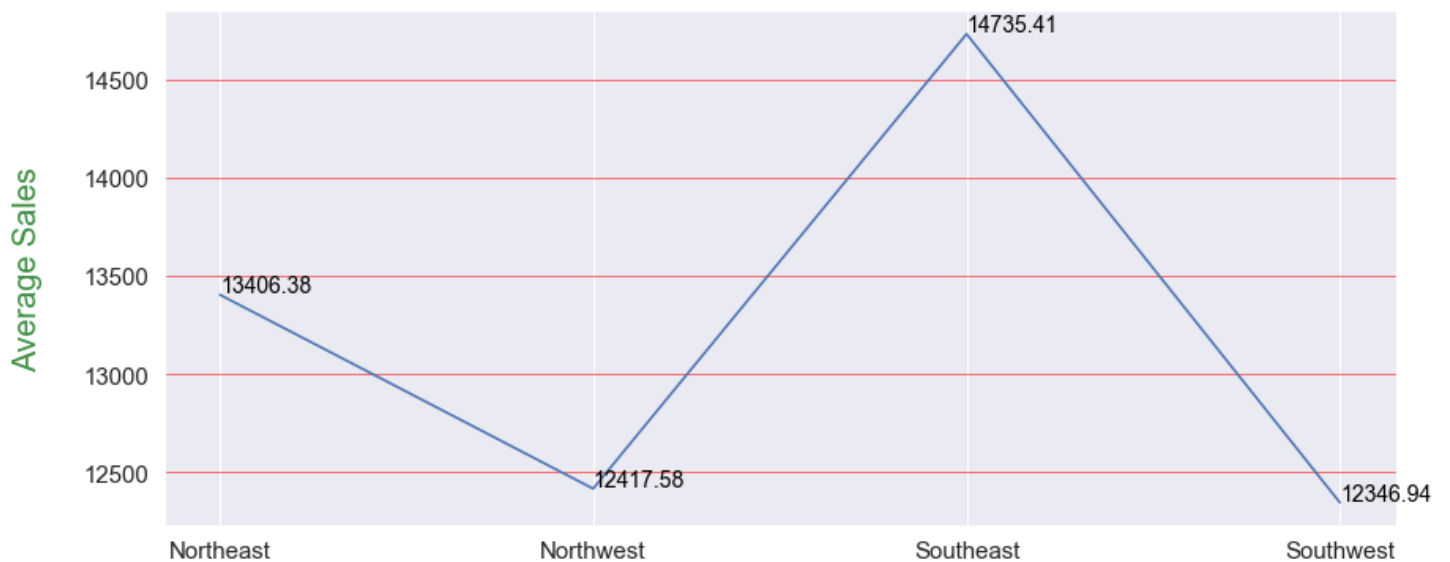


Smoker Percentage



Percentage of Smoker and Non Smoker

Average Sales per Region



Regions

Average Sales Per Region

in (2) internship in x what polynomia x Polynomial regi x Polynomial Regi x spyder replace x shortcut - How x Page Title x

File | C:/Users/hasij/OneDrive/Desktop/New%20folder/StartUp_Page.html

Apps ★ Bookmarks Known Sense Using the Android ... www.edudel.nic.in/... java - Changing Gri... History Imported From IE Google Portal Other bookmarks

Data Visualization

Data Visualization is a combination of science and art to display information and data in graphical format. By using components like charts, graphs, maps and data visualization tools provide an accessible information, patterns, outliers and relation between the data.

During the course CS503B and CS504B we learnt the fundamentals of R and Python. These fundamentals are further used to visualise the data present. This visualization helped us to understand the relation between the different elements of the data.

For this project I have visualized the data for an insurance organisation. Insurance industry is one of the backbone industry of any country. Using data of an insurance industry we will help this organization to visualize some data and relation between different element like Age, Gender, Body Mass Index(BMI), Number of Children, Smoker Status of Person, Region from which Person belongs and Cost of Insurance. This visualization will help to understand the trends and shape the future decisions.

Next Page

Type here to search

4:44 PM 2019-03-29

in (2) internship in x what polynomia x Polynomial regi x Polynomial Regi x spyder replace x shortcut - How x Page Title x

File | C:/Users/hasij/OneDrive/Desktop/New%20folder/Pie_chart.html

Apps ★ Bookmarks Known Sense Using the Android ... www.edudel.nic.in/... java - Changing Gri... History Imported From IE Google Portal Other bookmarks

Pie chart

A pie is a circular statistical graphic, which is divided into slices to illustrate numerical proportion. A chart with one or more sectors separated from the disk are known as exploded pie chart.



Gender	Percentage
Male	50.5%
Female	49.5%

This pie chart depicts the percentage of women and men considered in data.

Prev Page


Next Page

Type here to search

4:44 PM 2019-03-29

Bar Graph

A Bar graph is a chart or graph that is used to present the categorical data with rectangular bar with heights and lengths proportional to the values they are representing. Bars can be plotted vertically or horizontally.



Region	Number of people
Northeast	250
NorthWest	240
Southeast	320
Southwest	240


The above bar graph depicts the number of policy holders in each region.

Prev Page

Next Page

Scatter Plot

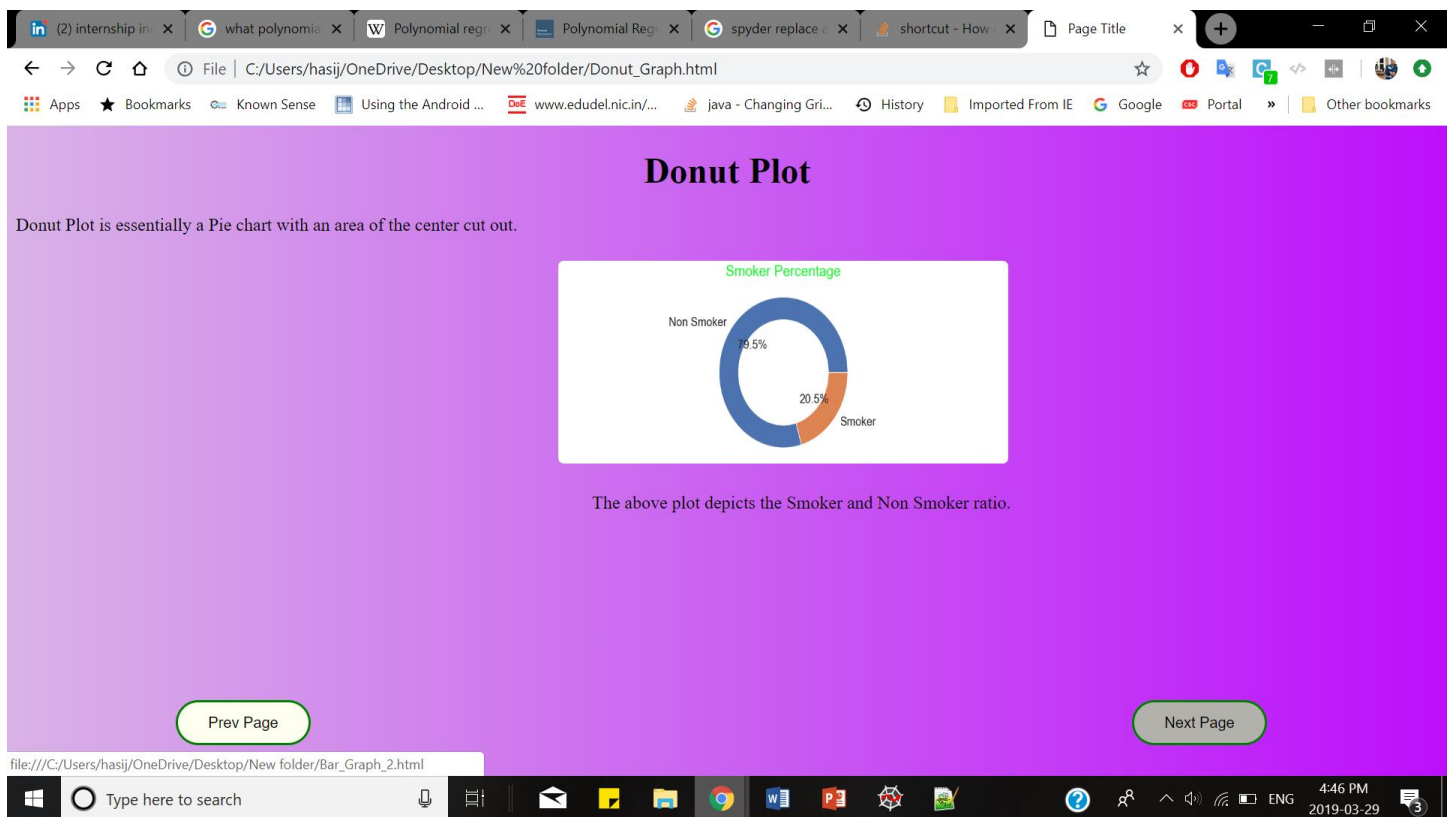
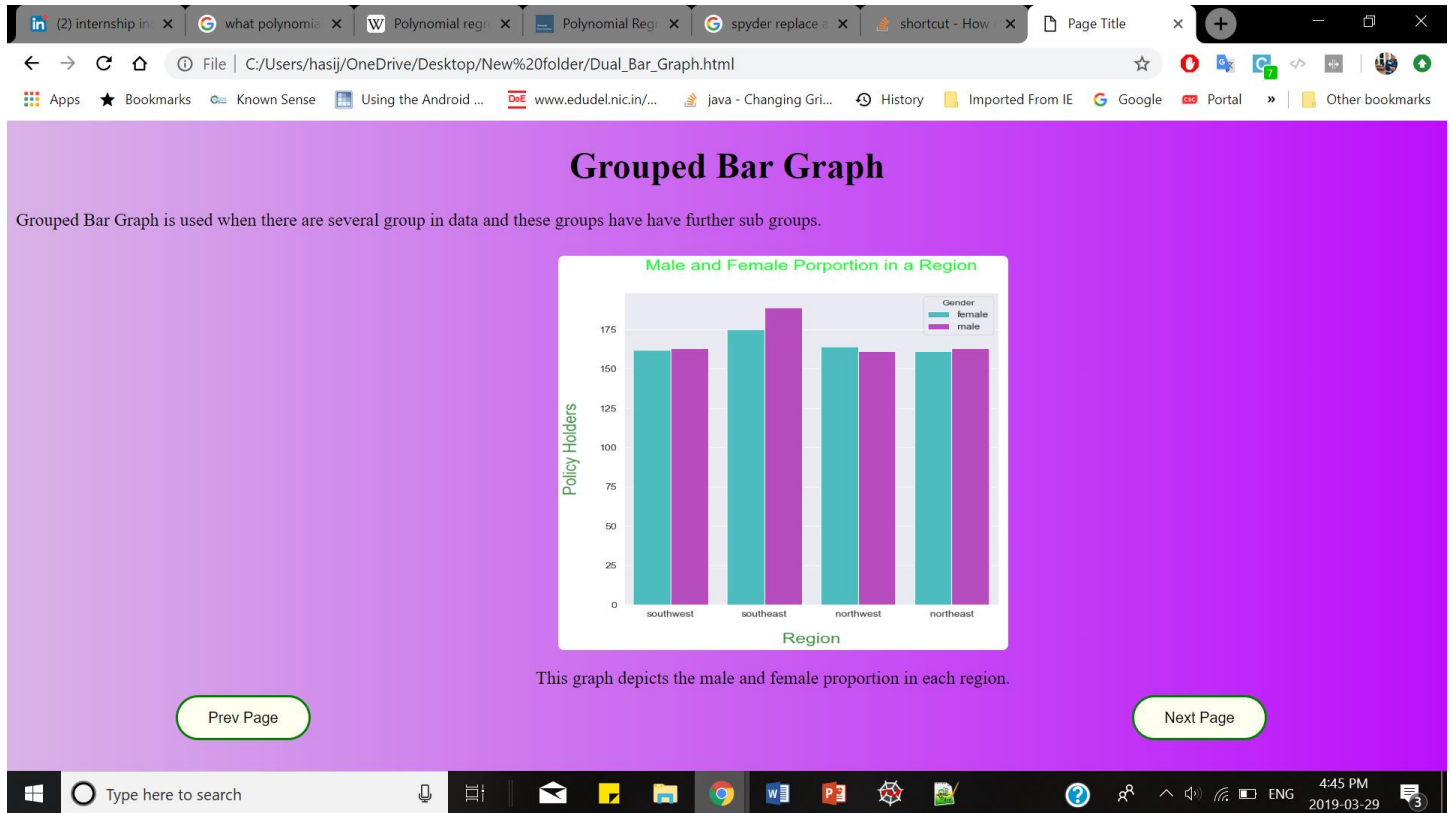
Scatter Plot is a type of plot that shows the collection of points. The position of points depends on its two-dimensional values.



This chart shows the Cost of Insurance for Smoker and Non Smoker on the basis of their Age.

Prev Page

Next Page



Bar Graph

A Bar graph is a chart or graph that is used to present the categorical data with rectangular bar with heights and lengths proportional to the values they are representing. Bars can be plotted vertically or horizontally.



BMI Group	Average Cost
low	8852.2
Normal	10404.9
high	13936.62

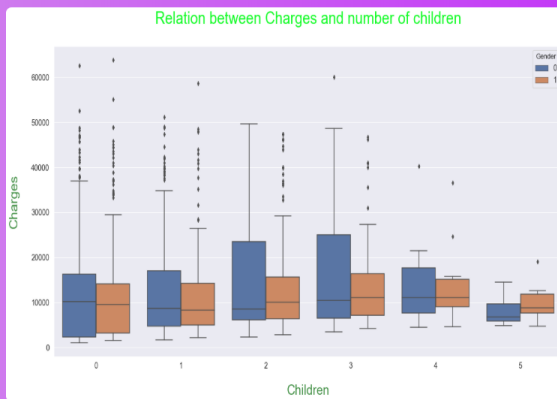
The above bar graph shows the average charge according to BMI group.

Prev Page

Next Page

Box Plot

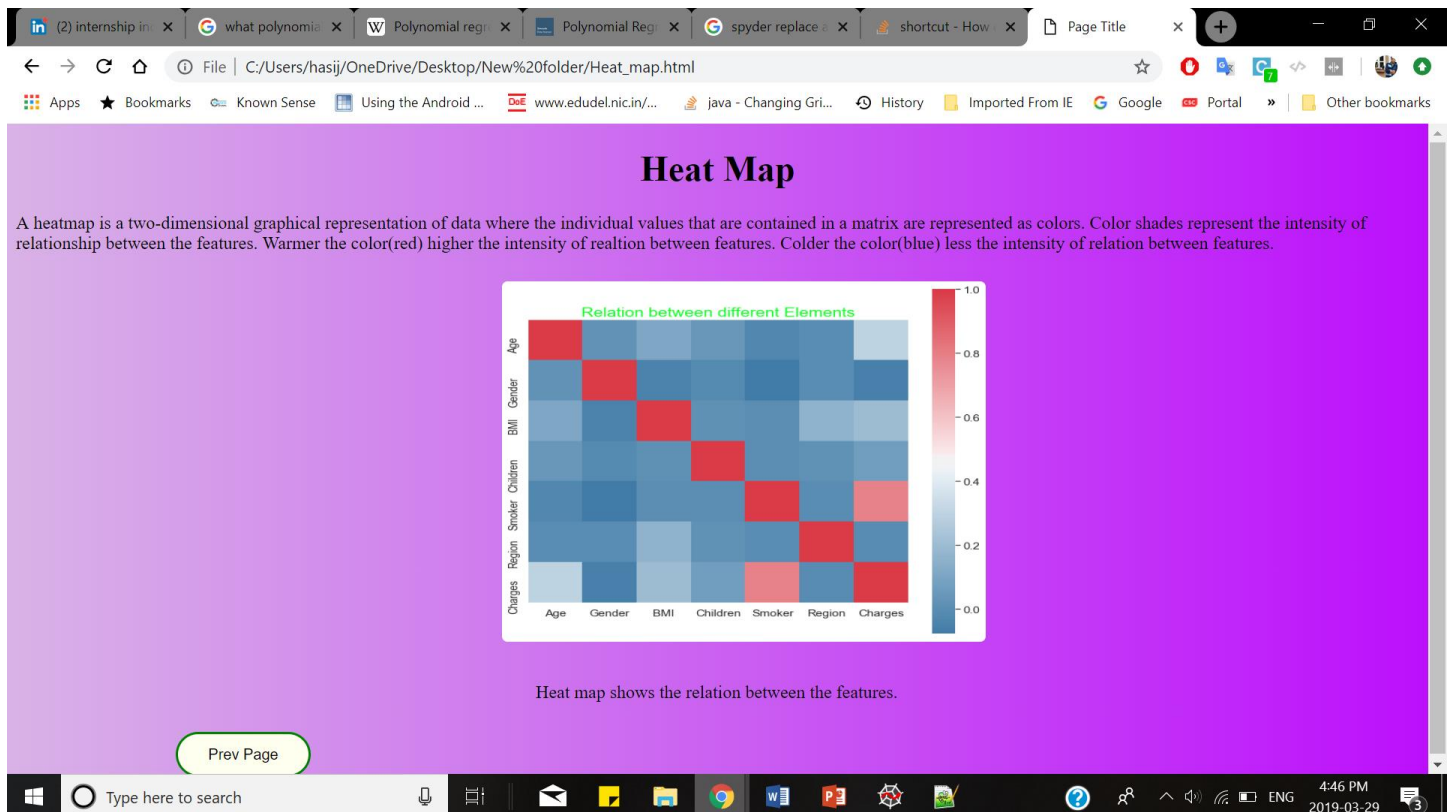
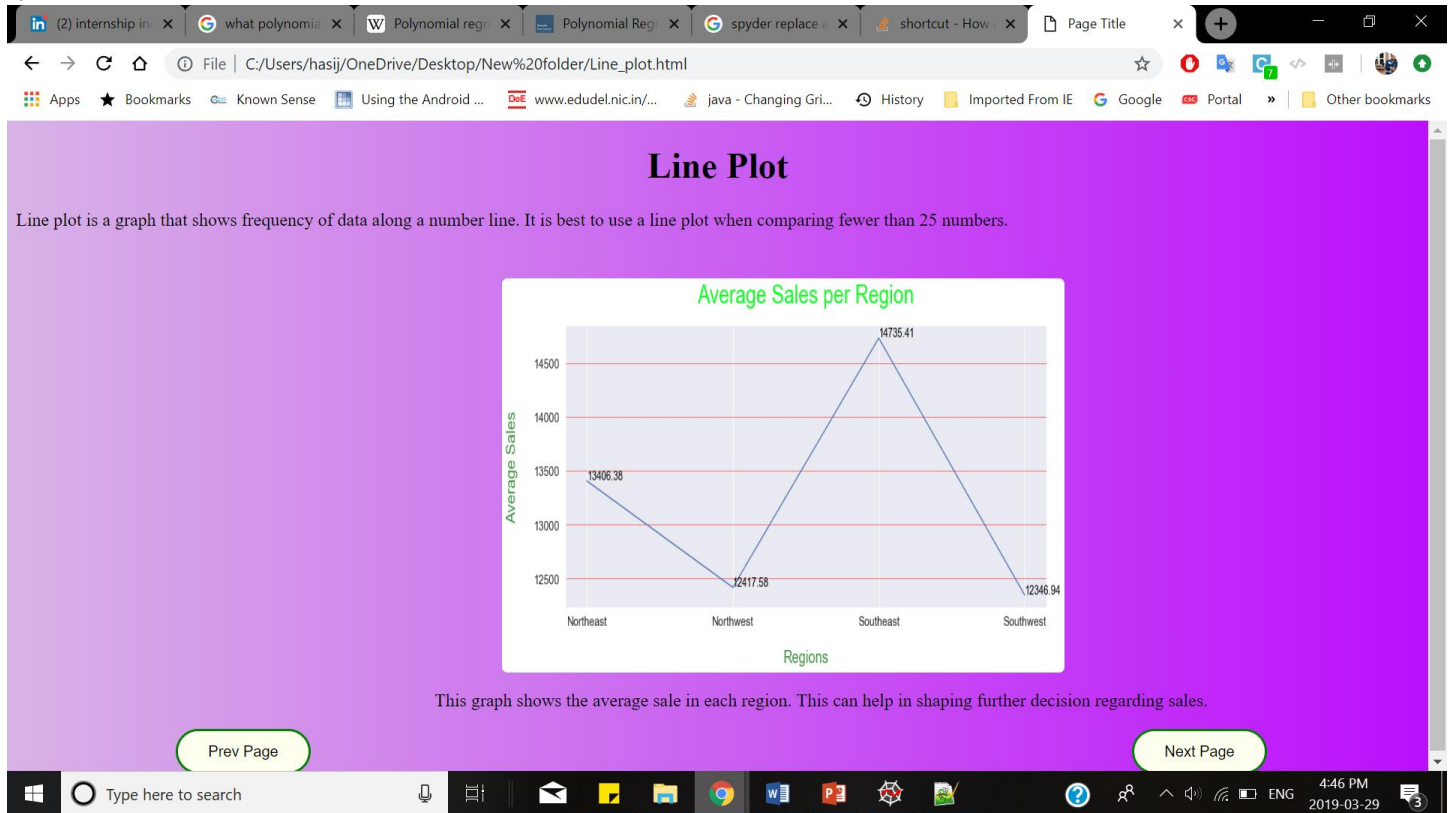
Box plot is a standardized way of displaying the distribution of data based on five number summary: minimum, first quartile, median, third quartile and maximum. In simple box plot central rectangular spans first quartile and third quartile. Middle segment of rectangular span is median. While above and below segments are maximum and minimum points. Points outside these are outliers.



Above graph shows the charge of male and female according to the number of children

Prev Page

Next Page



REFERENCES:

- [1] Machine learning Alogrithm: A review by Ayon Dev, Department of CSE, Gautam Budha University, Greater Noida , Uttar Pradesh, India
- [2] A Few Useful Things to Know about Machine learning, Pedro Dominque, Department of CSE, University of Washington, Seattle, USA
- [3] Random Forest Classifier from early developments to recent advancements, khaled Fawagreh, Mohamed Medhat Gaber,
- [4] Out-of-Bag Estimation, Leo Breiman, Statistics Department, University of California, Berkeley, USA
- [5] <https://blog.algorithmia.com/introduction-to-unsupervised-learning/>
- [6] <https://www.ngdata.com/what-is-data-analysis/>
- [7] https://ori.hhs.gov/education/products/n_illinois_u/datamanagement/datopic.html
- [8] <https://www.investopedia.com/terms/d/data-analytics.asp>
- [9] <https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/data-analysis/>
- [10] <https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/data-analysis/>
- [11] <https://www.tableau.com/learn/articles/data-visualization>