Technical Report Coversheet

Please complete all sections below and attach the completed coversheet to the front of your electronic assignment before submission:

Student Number: 199110906
Programme: Data Science Degree Apprenticeship
Module Tutor: Uche Onyekpe
Module Code: DSC5007M
Module Title: Applied Machine Learning
Report Title: Technical Report
Word Count: 2963

Declaration of Academic Integrity

Please complete before submitting your assignment:

By entering an ‘x’ in the box above, I confirm that I have read and understood the University regulations on cheating and plagiarism ‘ASS09 Cheating and Plagiarism’ and the work submitted is my own within the meaning of the regulations.

Applied Machine Learning

Machine learning algorithms for solving real-world regression, classification, and clustering problems

Linear Regression and Clustering analysis on Medical Cost Insurance Data

This case study describes how linear regression and clustering algorithms applied on the medical cost insurance data in machine learning.

Applied machine learning refers to ***the application of machine learning to address different data-related problems***.

Machine learning (**ML**) is a field of inquiry devoted to understanding and building methods that 'learn', that is, methods that leverage data to improve performance on some set of tasks.

It is found that the systems are analysing vast amounts of data over time to carry out tasks independently with ML resulting in better insights. ML solutions simply learn from experience without being programmed and this Artificial Intelligence (AI) has come a long way in the past decade and is now an integral part of our day to day lives.

ML can go through the thousands of customer communications per day and comprehend whether a customer is upset and making a complaint or whether a certain offer would make a customer happy. It can also make predictions, such as whether customers would like to purchase a specific item or would rather terminate services.

ML can be used to explore customers data, while learning how to predict spending trends or satisfaction with the company. It can provide personalized service to customers based on past interactions and make decisions and predictions with a degree of certainty while modifying its approach as it learns.

The purpose of ML is that computers can act more like human, with all the benefits of computing. Example, natural language processing. Computers can communicate with customers as naturally as a human customer service colleague.

The predicting medical insurance costs using ML approaches is still a problem in the healthcare industry that requires investigation and improvement. Using a series of ML algorithms, this study provides a computational intelligence approach for predicting healthcare insurance costs.

Insurance companies that sell life, health, and property and casualty insurance are using machine learning (ML) to drive improvements in customer service, fraud detection, and operational efficiency. For example, the Azure cloud is helping insurance brands save time and effort using machine learning to assess damage in accidents, identify anomalies in billing, and more.

We are going to be using various approaches using linear regression and K-means clustering.

Introduction:

The Insurance industry is rapidly growing and using machine learning/artificial intelligence techniques to enhance customer service, build better underwriting models, price prediction, claims handling, etc., It is doing so by leveraging the vast amount of data collected over the past few years. Data is the backbone of any machine learning technique and the insurance industry had been capturing data from various sources and using it for solving important business problems and accelerate business.

Health insurance companies today are using artificial intelligence (AI) and ML in ways not possible just five years ago to better pinpoint at-risk individuals and to reduce costs.
AI in medicine uses data science and algorithms to recognize patterns in medical data and then generate meaningful predictions and outputs.

ML allows building models to quickly analyse data and deliver results, leveraging historical and real-time data. With ML, healthcare service providers can make better decisions on patient's diagnoses and treatment options, which lead to an overall improvement of healthcare services.

To offer competitive rates, the insurance industry needs predictability. That's why insurers are increasingly turning to AI and ML to more readily predict the future and the risks that the future may hold for their customers.

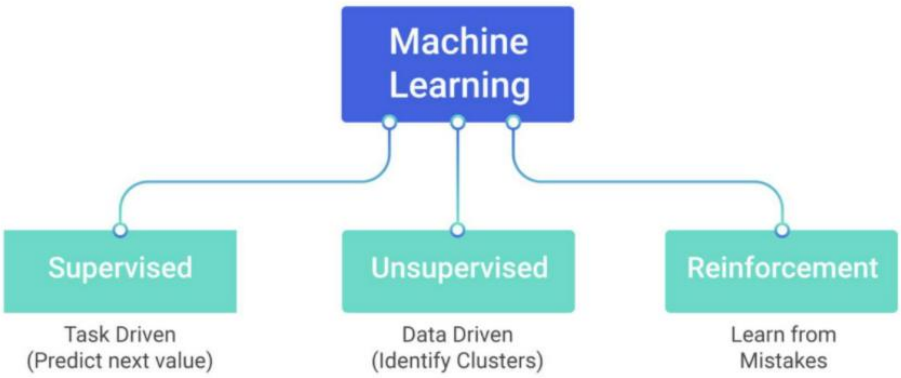
For example, AI can help insurance companies to gain a sufficient insight into changing weather patterns to keep premiums low, by analysing historical weather data and future climate predictions.

AI allows insurance companies to rely on predictions based on real events in near-real time using large datasets, as opposed to statistical sampling of past performance.

It can also help financial firms to stay compliant with the ever-changing regulations to which they are subject. Machine learning algorithms can quickly read and learn from regulatory documents to detect correlation between certain actions and compliance, which in turn enables the detection of anomalies. Machine learning algorithms can be used in the prevention of fraudulent claims, too, by identifying characteristics that set them apart from legitimate claims.

ML can help to improve fraud detection, enhance customer service, and reduce expenditures in the health insurance sector.

Machine Learning is classified into 3 types:



Types of Machine Learning

Supervised Learning algorithms are used when we have labelled data and are trying to predict a label (target) based off of known features (input variables).

This is commonly used in applications where historical data predicts likely future events. For example, it can attempt to predict the price of a product/car/house based on different features for products for which we have historical price data.

Unsupervised Learning algorithms are used when we have unlabelled data and are trying to group together similar data points based off features.

This is mainly used to explore the data and find some structure within. For example, it can identify the image (cat or dog) based on different inputs which groups together similar segments and then attempts to recognize the image correctly. This is unsupervised learning, where a machine is not taught but learns from the data (in this case data about a dog or cat).

Reinforcement Learning occurs when a computer system receives data in a specific environment and then learns how to maximize its outcomes. That means this model keeps continues to learn until best possible behaviour is met. Reinforcement learning is frequently used for robotics, gaming, and navigation.

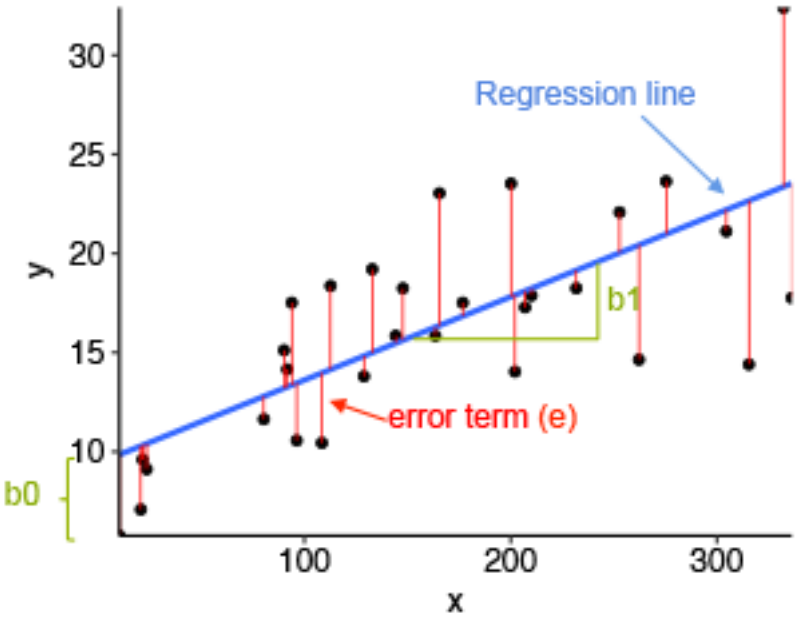
Information gathering and processing systems are being updated and streamlined for efficacy. Machine learning can be applied in healthcare through **lowering the cost and chaos of recordkeeping, including electronic health records, and maintaining data integrity**

Linear Regression Algorithm

Linear regression is one of the most used techniques in statistics. It is used to quantify the relationship between one or more predictor variables and a response variable.

The most basic form of linear is regression is known as **simple linear regression,** which is used to quantify the relationship between one predictor variable and one response variable. Linear Regression is the **first machine learning algorithm** based on '**Supervised Learning'**. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x).

If we have more than one predictor variable then we can use **multiple linear regression**, which is used to quantify the relationship between several predictor variables and a response variable.



Simple linear regression equation, where b_o is the intercept, b_1 is coefficient or slope, x is the independent variable and y is the dependent variable.

$$y = b_o + b_1 x$$

Multiple linear regression equation, where bo is the intercept, $b_1, b_2, b_3, b_4 \ldots, b_n$ are coefficients or slopes of the independent variables $x_1, x_2, x_3, x_4 \ldots, x_n$ and y is the dependent variable.

$$y = b_o + b_1 x_1 + b_2 x_2 + b_3 x_3 \ldots . + b_n x_n$$

In the above diagram:

- x is our dependent variable which is plotted on the x-axis and y is the dependent variable which is plotted on the y-axis.
- Black dots are the data points i.e the actual values.
- b_o is the intercept which is 10 and b_1 is the slope of the x variable.
- The blue line is the best fit line predicted by the model i.e the predicted values lie on the blue line.
-

The vertical distance between the data point and the regression line is known as **error or residual.**
Each data point has one residual and the sum of all the differences is known as **the Sum of Residuals/Errors.**

Mathematical Approach:

Residual/Error = Actual values – Predicted Values

Sum of Residuals/Errors = Sum(Actual- Predicted Values)

Square of Sum of Residuals/Errors = (Sum(Actual- Predicted Values))^2

i.e

$$\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$$

Regression analysis is an important statistical method for the analysis of medical data. It enables the identification and characterization of relationships among multiple factors. Medical researchers often use linear regression to understand the relationship between drug dosage and blood pressure of patients.

In general, the regression analysis is helpful for the decision-making process within the organisation due to several reasons. The first point is that the description of the relationship between the variables helps to establish the existence of a possible causal relationship. The second one is that compiling the regression can predict the values of the dependent variable from the values of the independent variables, which allows determining the predictor for the dependent variable.

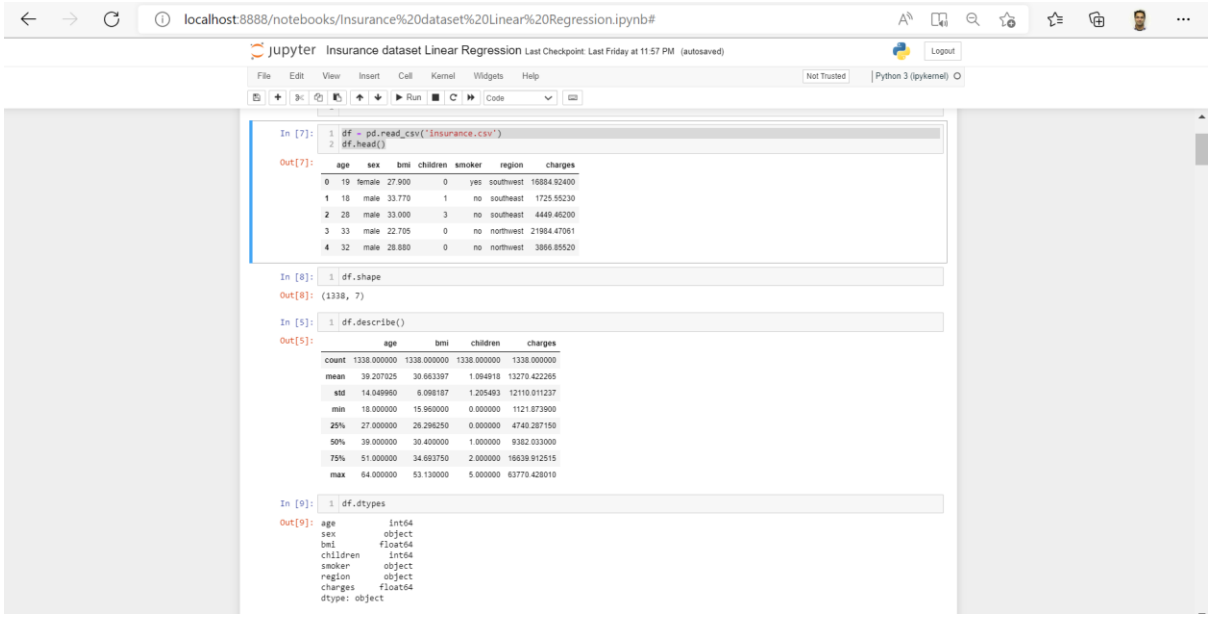
In this case study, there are factors that affect how much health insurance premiums cost:

1) **age:** Age of primary beneficiary
2) **sex:** Insurance contractor gender, male or female
3) **bmi:** Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body we/ight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
4) **children:** Number of children covered by health insurance / Number of dependents
5) **smoker:** Smoking
6) **region:** the beneficiary's residential area in the US, northeast, southeast, southwest, northwest

We have started off with the exploratory data analysis before continuing with using linear regression application as well as clustering.

Before doing data analysis, it is important to prepare, clean and explore the data to better help you reach quality insights.

It is always important to check the types of the labels so we will know how we would handle the data based on its kind, such as int64, object or float64.



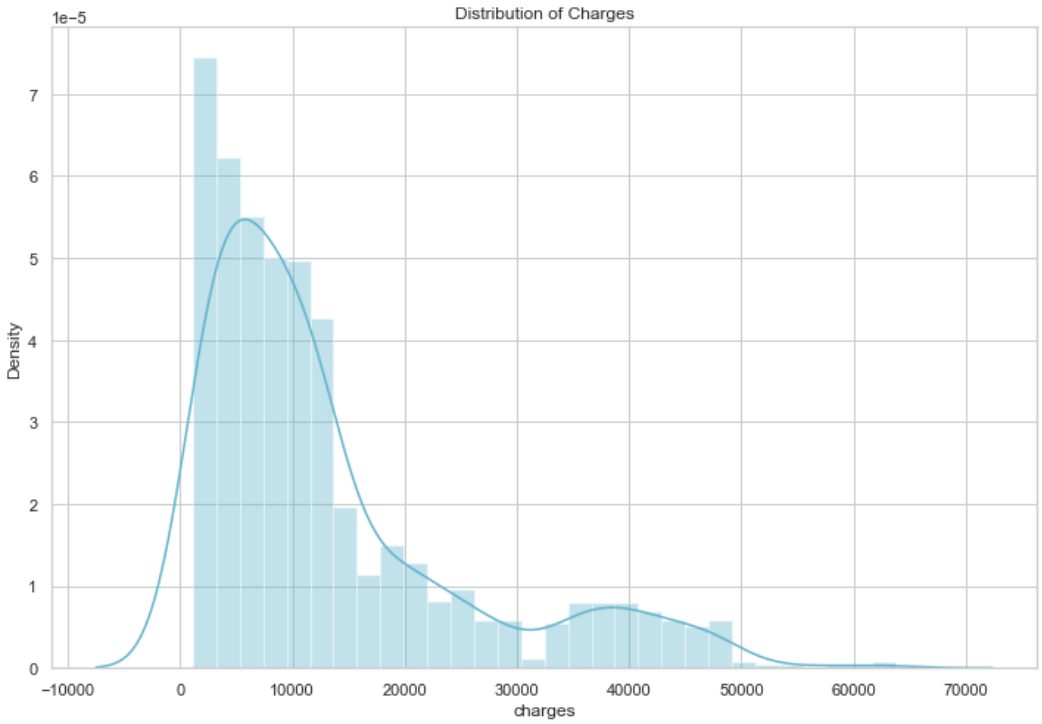

```
In [9]:  1 df.dtypes

Out[9]: age           int64
        sex          object
        bmi         float64
        children      int64
        smoker       object
        region       object
        charges     float64
        dtype: object
```

We also check to make sure that there are no null values by entering df.isnull().sum(). This shows that we have zero values which is very good.

```
Out[10]: age         0
         sex         0
         bmi         0
         children    0
         smoker      0
         region      0
         charges     0
         dtype: int64
```

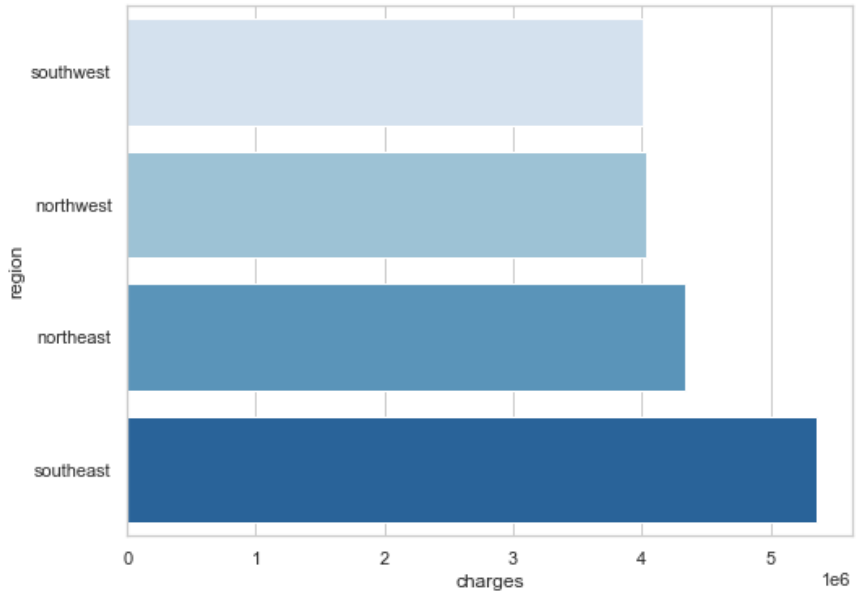
How charges are affected based on the given factors.

This distribution is right-skewed. To make it closer to normal we can apply natural log.



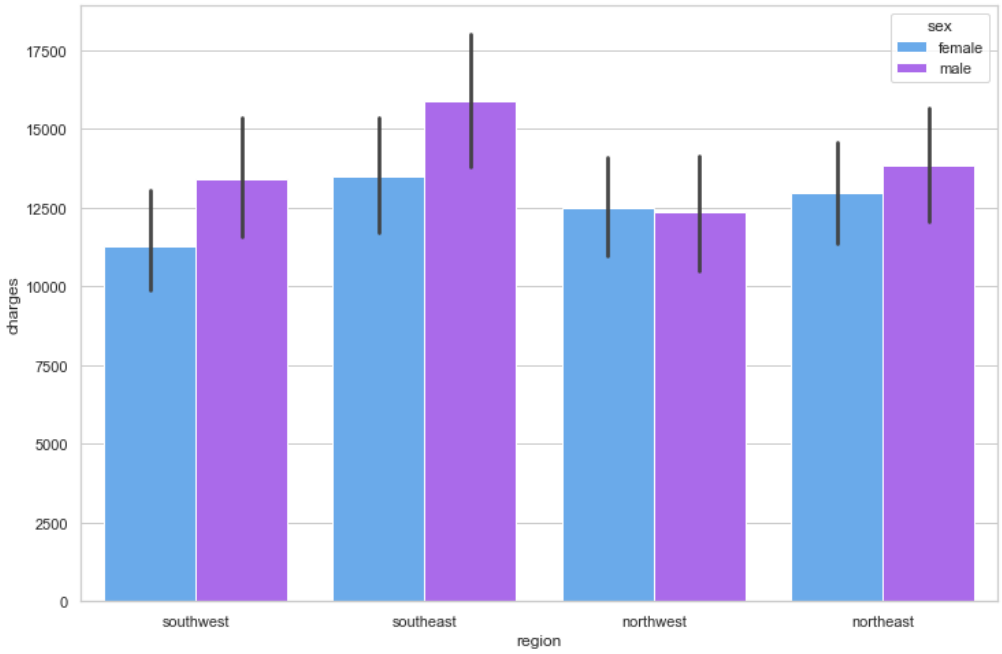
Right skewed distribution is also called as positive skewed distribution. This is a type of distribution in which most values are clustered around the left tail of the distribution while the right tail of the distribution is longer.

Charges by Region

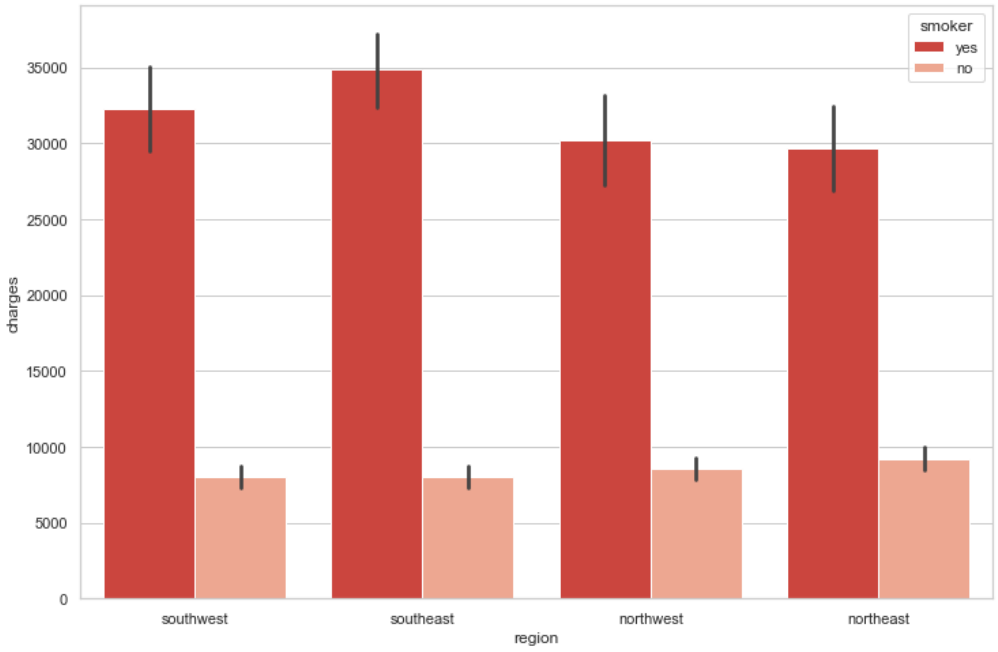


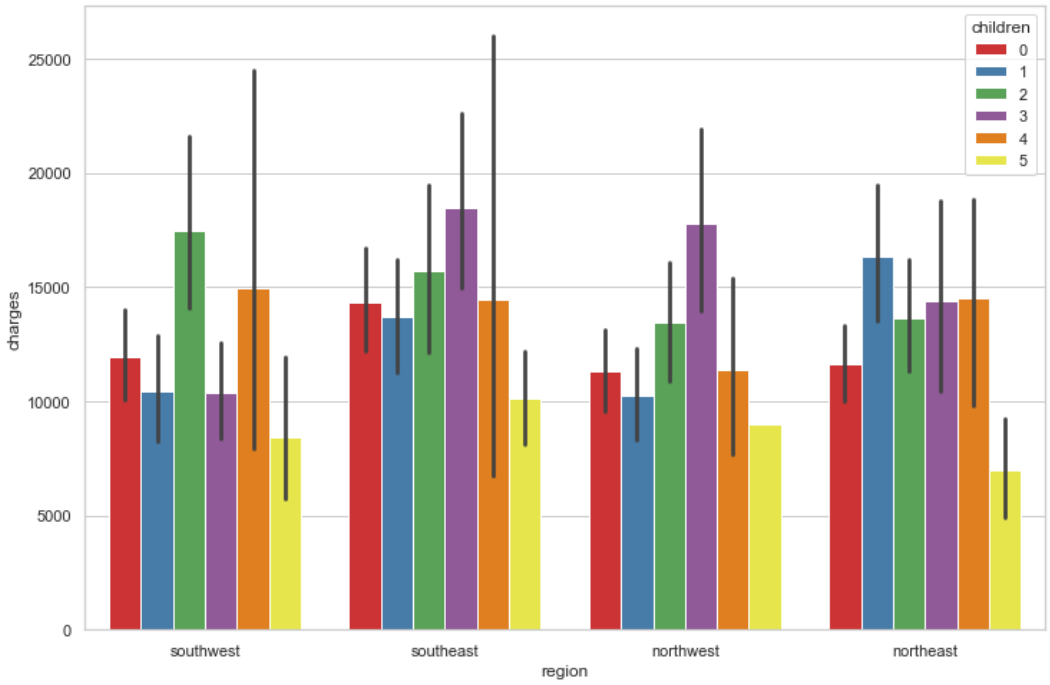
The highest medical charges are in the Southeast and the lowest seems to be in the Southwest.

Considering certain factors (sex, smoking, having children) let's see how it changes by region



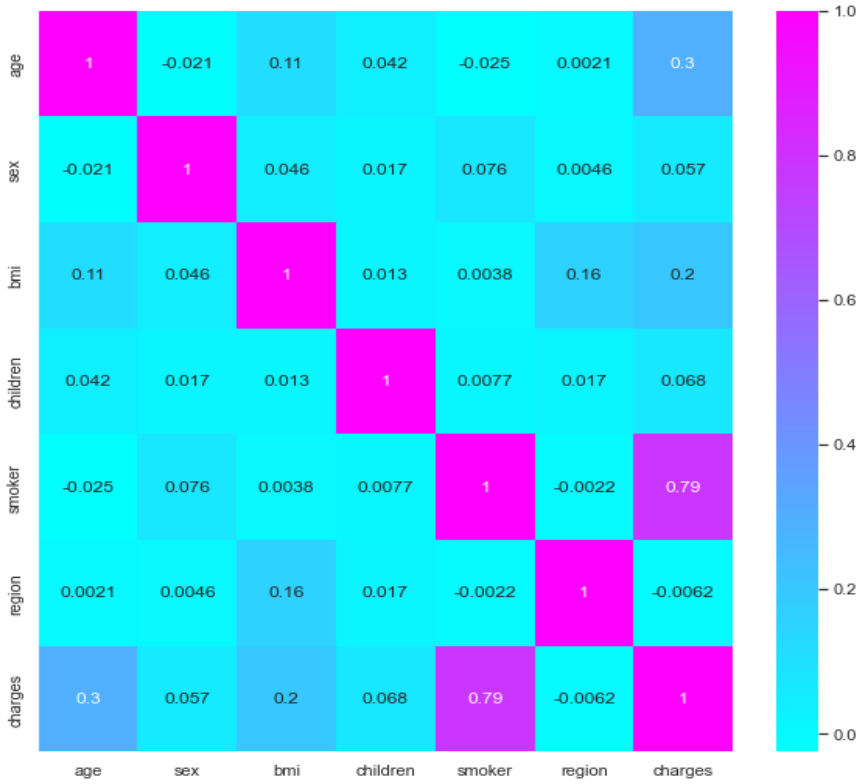
We can see from these bar graph Southeast region seems to be the one that is impacted by the increase in charges and are the highest across all other regions.
but the lowest are in the Northwest.





We are also considering and analysing performance of the people who either smoke or don't smoke and with or without children.

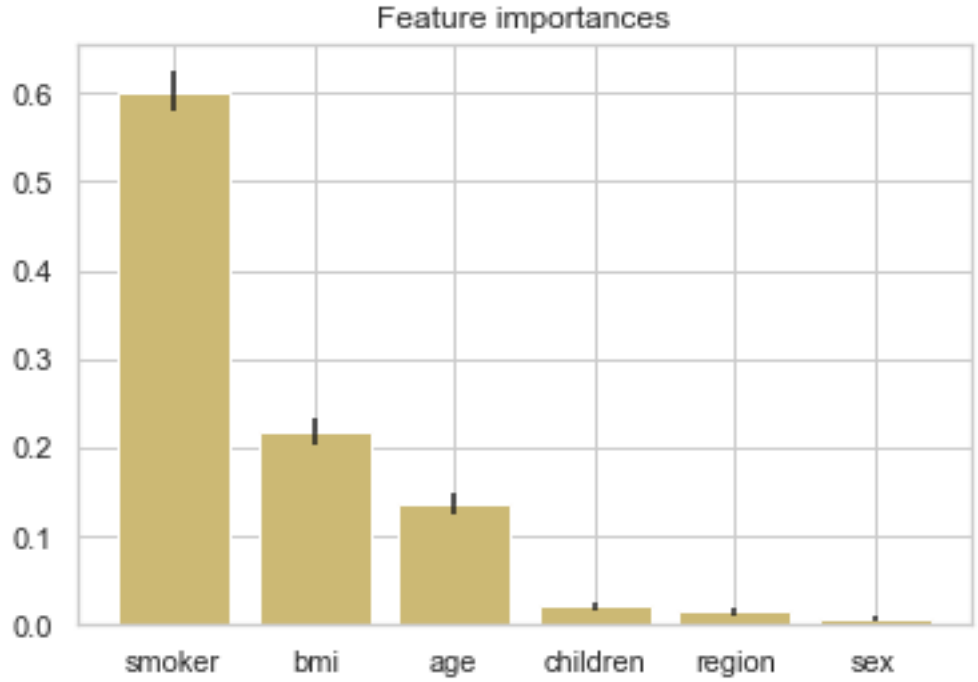
People in the Southeast generally smoke more than people in the North east and North west.



It is also identified that the people in the Northeast have higher charges by gender than in the Southwest and Northwest overall. And people with children tend to have higher medical costs overall as well.

There is no correlation, except with the smoking.

The result we got from the above visualisation is good enough, but we can try to improve it a bit by reducing unimportant features later.



We have tried to apply different regression methods such as Linear, Ridge, Lasso, Random Forest Regressor and polynomial regression.

Smoking is the greatest factor that affects medical cost charges, then it's bmi and age. Polynomial Regression turned out to be the best model.

Clustering Analysis

Cluster analysis or clustering is an **unsupervised machine learning algorithm** that groups unlabelled datasets. It aims to form clusters or groups using the data points in a dataset in such a way that there is high intra-cluster similarity and low inter-cluster similarity. In, layman terms clustering aims at forming subsets or groups within a dataset consisting of data points which are really like each other and the groups or subsets or clusters formed can be significantly differentiated from each other.

Different types of Clustering Algorithms:

a) **K-means Clustering** – Using this algorithm, we classify a given data set through a certain number of predetermined clusters or “k” clusters.

b) **Hierarchical Clustering** – follows two approaches Divisive and Agglomerative.

c) Agglomerative considers each observation as a single cluster then grouping similar data points until fused into a single cluster and Divisive works just opposite to it.

d) **Fuzzy C means Clustering** – The working of the FCM Algorithm is almost similar to the k-means clustering algorithm, the major difference is that in FCM a data point can be put into more than one cluster.

e) **Density-Based Spatial Clustering** – Useful in the application areas where we require non-linear cluster structures, purely based on density.

In this case study, we will be looking at K-means clustering.

K-Means Clustering:
K-Means Clustering is an Unsupervised Learning algorithm, used to group the unlabeled dataset into different clusters/subsets.

K’ defines the number of pre-defined clusters that need to be created in the process of clustering say if k=2, there will be two clusters, and for k=3, there will be three clusters, and so on. As it is a centroid-based algorithm, ‘means’ in k-means clustering is related to the centroid of data points where each cluster is associated with a centroid.

k-means clustering algorithm performs two tasks:

- Determines the most optimal value for K center points or centroids by a repetitive process.
- Assigns each data point to its closest k-center. Cluster is created with data points which are near to the particular k-center.

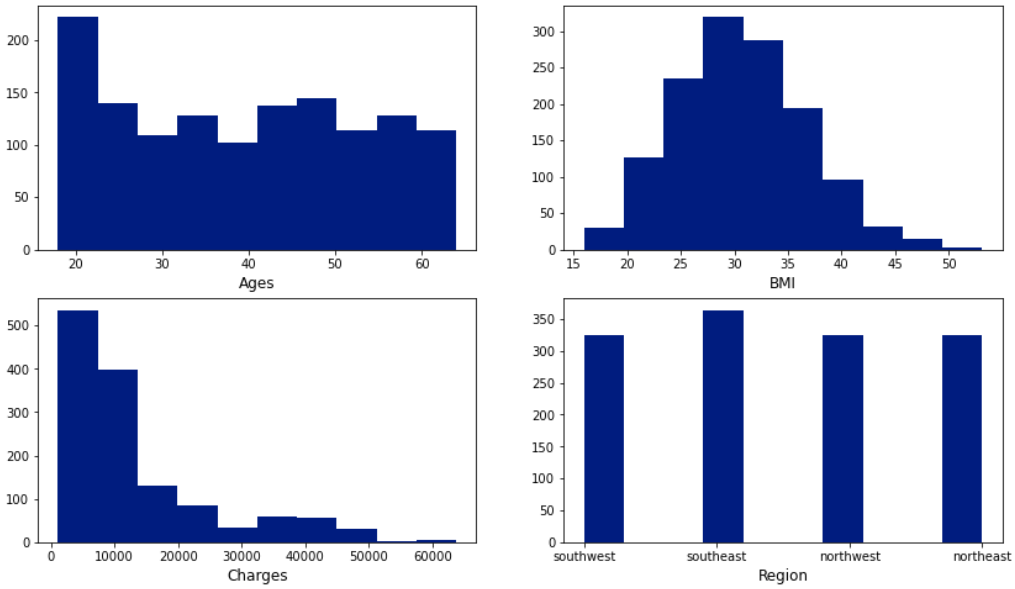
We start this clustering analysis by looking at the basic histogram. This helps us to find the type of data we will be using for analysis.

The basic histogram, called by the hist() function gives a visual represention of the distribution of the data based on the Normal distribution. The graphs below shows:

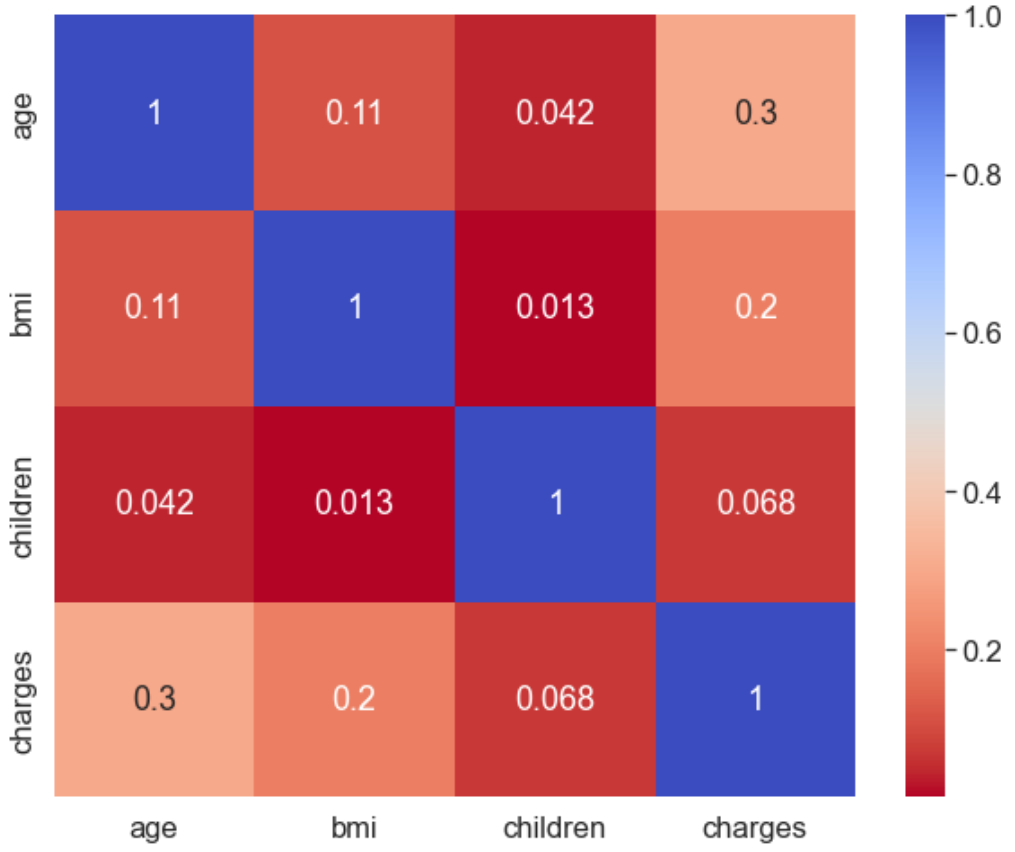
- Data type of each variable: "Continuous variables" will have a continuous normal distribution curve. "Categorical variable" will have distinct plots.

- Skewness of data highlights the presence of outliers. Here we note Charges and Age have outliers where as BMI doesn't.

- Region is categorical variable.



Correlation:
Correlation is a statistical term describing the degree to which two variables move in coordination with one another. If the two variables move in the same direction, then those



variables are said to have a positive correlation. If they move in opposite directions, then they have a negative correlation.

We can see age and charges have very slight positive correlation with charges which we will try to prove in due course.

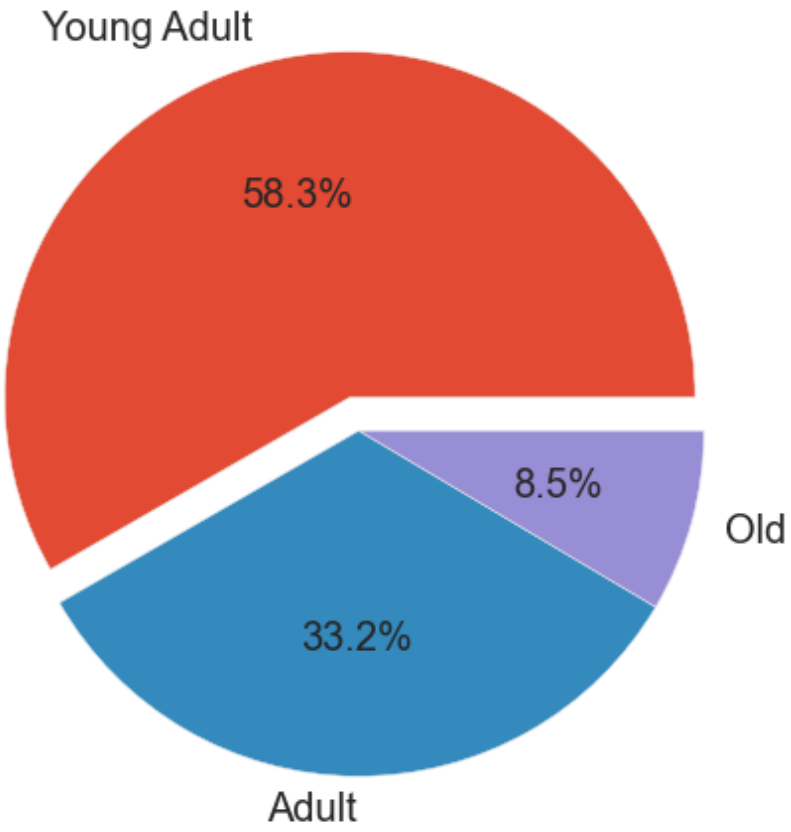
Data shows some interesting facts about the relationship between age and charges.
We will first convert the ‘age’ into bins/groups of categorical variables like Child, Young Adult, Adult and Old to analyse its relationship with medical expenses "charges".

	age	sex	bmi	children	smoker	region	charges	age_cat
0	19	female	27.900	0	yes	southwest	16884.92400	Young Adult
1	18	male	33.770	1	no	southeast	1725.55230	Young Adult
2	28	male	33.000	3	no	southeast	4449.46200	Young Adult
3	33	male	22.705	0	no	northwest	21984.47061	Adult
4	32	male	28.880	0	no	northwest	3866.85520	Adult
...	...	...	...	...	...	...	...	...
1333	50	male	30.970	3	no	northwest	10600.54830	Adult
1334	18	female	31.920	0	no	northeast	2205.98080	Young Adult
1335	18	female	36.850	0	no	southeast	1629.83350	Young Adult
1336	21	female	25.800	0	no	southwest	2007.94500	Young Adult
1337	61	female	29.070	0	yes	northwest	29141.36030	Old

1338 rows × 8 columns

An illustration below using pie chart shows the large population of age group are young adults (58.3%) to adults (33.2%).

It means that there is significant proportion of young adults smoking as well.



From the above three bar plots we note the following:-

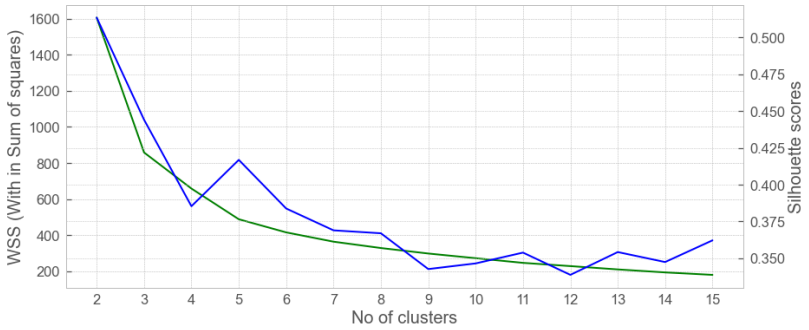
Adults comprises of 33.2% of the whole data set and thus sum total of their medical expenses is the highest but the mean cost per adult patient is less than $15,000 with a standard deviation of $12,000. Adult age ranges from 30 to 59 with critical age starts post 59 where lots of ailments crop up due to work stress and socio-environmental factors.

Young Adult is the age ranges from 18 to 30. A age when human body is at its peak. With a 58.3% of Young adults representation still get lowest total. With a mean of around less than $10,000 and standard deviation of around $10,000

Old age is the age where the medical cost becomes the primary expenditure and its evident by the fact that the mean cost is among the highest which shoots above $20,000 with a standard deviation of $13,000.

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

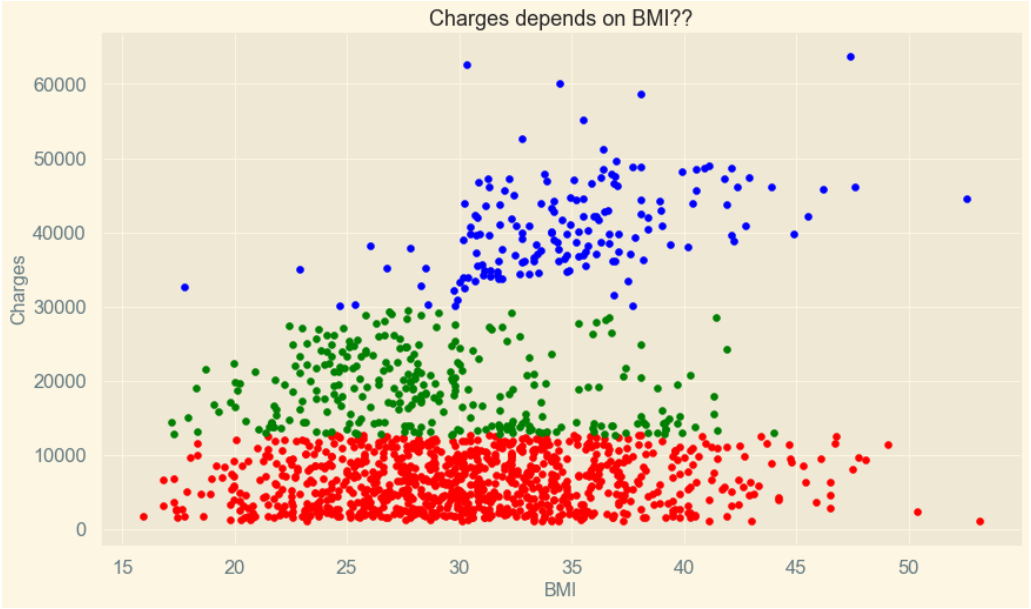
Clustering is a Machine Learning technique whose aim is to group the data points having similar properties and/or features, while data points in different groups should have highly offbeat properties and/or features.



The graph shows that we see the "elbow" at 3 and silhouutee score almost best at that point. We are then able to define the clusters as shown in the table below:

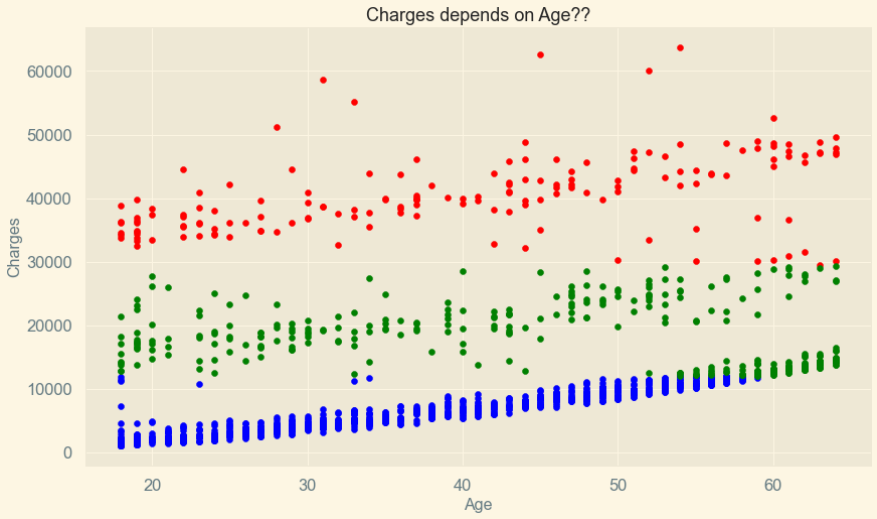
	age	sex	bmi	children	smoker	region	charges	age_cat	log_charges	bmi_cat	clusters
1337	61	female	29.070	0	yes	northwest	29141.36030	Old	10.279914	Obese	A
1020	51	male	37.000	0	no	southwest	8798.59300	Adult	9.082347	Obese	A
489	53	male	31.160	1	no	northwest	10461.97940	Adult	9.255503	Obese	A
491	61	female	25.080	0	no	southeast	24513.09126	Old	10.106963	Obese	A
1018	54	female	35.815	3	no	northwest	12495.29085	Adult	9.433107	Obese	A
...	...	...	...	...	...	...	...	...	...	...	...
621	37	male	34.100	4	yes	southwest	40182.24600	Adult	10.601181	Obese	C
252	54	male	34.210	2	yes	southeast	44260.74990	Adult	10.697854	Obese	C
251	63	female	32.200	2	yes	southwest	47305.30500	Old	10.764378	Obese	C
271	50	male	34.200	2	yes	southwest	42856.83800	Adult	10.665620	Obese	C
668	62	male	32.015	0	yes	northeast	45710.20785	Old	10.730077	Obese	C

1338 rows × 11 columns



From the above as we have defined, we got 3 distinct clusters. With BMI (15 to 35) has a expense of $10,000 to $30,000 whereas higher BMI's have much higher cost and lower BMI will have much lower cost.

We don’t see much distinction about the age groups. All the three expenses ranges have all the age groups.



We can see that the charges are dependent on the above-mentioned variables.

Machine Learning can **help insurers assess risk, detect fraud and reduce human error in the application process**. The result is insurers who are better equipped to sell customers the plans most suited for them. Customers benefit from the streamlined service and claims processing that AI affords.

Python Codes for Linear Regression and Clustering (Python and Jupyter Notebook codes):

Insurance dataset Linear Regression.ipynb

Insurance dataset Clustering.ipynb

Insurance dataset Linear Regression.py

Insurance dataset Clustering.py