Report on
Project Assignment on Python

Course Name: Introduction to Data Science with Python

Course Code: WM-ASDS04 4/21/23 Submission Date: 4/21/23

Submitted by:

S.N.	Name	Roll
01	Md. Nazmul Huda	20229033
02	S M Asaduzzaman	20229017
03	Asad All Farabi	20229013
04	Md Amanullah Rishat	20229010
05	Rafsan Zaman	20229016

Section-A

Semester-1st, Batch-9th

Submitted to:

Farhana Afrin Duty
Assistant Professor
Department of Statistics
Jahangirnagar University

জাহাঙ্গীরনগর বিশ্ববিদ্যালয়

Weekend Masters in Applied Statistics and Data Science
Department of Statistics
Jahangirnagar University
Savar, Dhaka-1342

Contents

1. Introduction 2

2. Problem statement 2

3. Objectives 3

4. Methodology 3

5. Findings 7

Output-1: Reference#6 7

Output-2: Reference#8 8

Output-3: Reference#9 9

Output-5: Reference #11 11

Output-6: Reference #12 12

Output-7: Reference #17 13

Output-8: Reference #18 & 18 13

6. Conclusion: 14

1. Introduction

The assignment: Final assignment or project report will be presented as part of Professional Master course on Introduction to Data science with python. The assignment will be completing by required steps suggested by faculty: 1. Select any one machine learning task from regression/classification Problem to solve; 2. Obtain the dataset; 3. Perform EDA for better understanding your datasets; 4. Data preprocessing to clean and make ready your data for analysis, such as handling missing value, scaling, encoding categorical variable etc.; 5. Build your Machine Learning Model using training data; 6. Evaluate performance of your model for test set and 7. Get your final model and prediction using the model.. The main objective of the assignment is to practice of sampling and statistical methods on hands after theoretical courses. Based on mentioned tasks a formal but short report need to be submitted.

2. Problem statement

The use of insurance data is crucial for insurance companies as it allows them to better understand their customers and develop effective pricing strategies. One of the key variables that is typically included in insurance data is demographic information, such as age, sex, BMI, and number of children. Additionally, lifestyle factors such as smoking habits can also have a significant impact on insurance charges.

Despite the importance of this data, there are several challenges associated with analyzing insurance data. One of the primary challenges is the complexity of the data, as it often contains numerous variables that are interdependent and difficult to analyze. Another challenge is ensuring that the data is accurate and reliable, as errors or inconsistencies in the data can lead to inaccurate conclusions and decisions.

Moreover, while insurance companies often use this data to develop pricing models and risk assessments, there is also concern about potential discrimination based on demographic or lifestyle factors. For example, charging higher insurance premiums to individuals based on their age, sex, or smoking habits could be considered discriminatory and lead to legal and ethical issues.

Therefore, there is a need for research to explore the potential biases and ethical considerations associated with using demographic and lifestyle data in insurance pricing models. Additionally, there is a need to develop more effective analytical techniques to better understand the complex relationships between these variables and insurance charges. By addressing these challenges, insurance companies can better leverage their data to develop more accurate and fair pricing strategies that benefit both the company and its customers.

3. Objectives

The main objective of this assignment is to apply machine learning in python techniques to analyze insurance data and build a model to predict insurance charges based on demographic and lifestyle factors. Specific objectives of the assignment are follows but not limited to:

i. To gain practical experience in data analysis, data preprocessing, and building machine learning models in python by the group of students.

ii. To develop critical thinking skills of the students by exploring potential biases and ethical considerations associated with using demographic and lifestyle data in insurance pricing models.

iii. To produce a formal report that summarizes the findings and conclusions of the analysis and feasibility of the model analyses by the group of students.

4. Methodology

To begin our analysis, we first gather open source data on insurance pricing and save it in CSV format. Next, we upload the CSV file to Google Drive and import the data into a Jupyter notebook using Python. Using the appropriate code, we perform the necessary analysis to obtain the desired outcome. List of code are as follows and in findings section we use the SL# as reference code to make better understanding of our analysis.

Table-1: Code used in analysis

SL# Reference	Applied Code

1.	from google.colab import drive
	drive.mount('/content/drive')

2.	%cd 'drive/MyDrive/Lab_Project/'
3.	#importing library import pandas as pd import numpy as np

4.	#reading the csv file
	data=pd.read_csv('data/insurance.csv')

5.	import seaborn as sns
	import matplotlib.pyplot as plt
	sns.heatmap(data.isnull(),yticklabels=False,cbar=False,cmap='Blues')

6.	plt.figure(figsize=(20,10))
	sns.set_style('dark')
	fig1=sns.countplot(x='age',data=data,palette='BrBG')
	plt.savefig('fig1.png',dpi=100)

SL# Reference	**Applied Code**

7.	from sklearn import preprocessing
	label_encoder=preprocessing.LabelEncoder()
	data1=data.copy()
	data1['smoker']=label_encoder.fit_transform(data['smoker'])

8.	pearson =data1.corr()
	fig, ax = plt.subplots(figsize=(10,7))
	fig2=sns.heatmap(data = pearson, vmin=-0.2, vmax=1, cmap= 'Blues_r',annot=True, fmt=".1g")
	plt.savefig('fig2.png',dpi=100)

9.	sfig, ax=plt.subplots(2,1, sharey=True, figsize=(16,12))
	ax[0].set_title('Distribution of age')
	fig3=sns.countplot(x='age', data=data, color='c', ax=ax[0])
	plt.savefig('fig3.png',dpi=100)
	ax[1].set_title('Distribution of age and smoker')
	fig4=sns.countplot(x='age',hue='smoker', data=data, palette='Blues',ax=ax[1])
	plt.savefig('fig4.png',dpi=100)

10.	fig, ax=plt.subplots(figsize=(15, 7))
	ax.set_title('1. Bar distribution of charges for age')
	fig5=sns.barplot(x='age',y='charges',data=data, color='c', ax=ax)
	plt.savefig('fig5.png',dpi=100)
	fig, ax=plt.subplots(figsize=(15, 7))
	ax.set_title('2. Scatter distribution of charges for age')
	fig6=sns.swarmplot(x='age', y='charges',hue='smoker',palette='Blues_r', data=data, ax=ax)
	plt.savefig('fig6.png',dpi=100)
	fig7=sns.lmplot(x='age', y='charges',hue='smoker',palette='Blues_r', data=data, aspect=2.5)
	plt.savefig('fig7.png',dpi=100)

11.	ig, ax = plt.subplots(1,3, sharey=True, figsize=(15,5))
	fig8=sns.countplot(x='region', data=data, ax=ax[0])
	plt.savefig('fig8.png',dpi=100)
	fig9=sns.countplot(x='region',hue='sex', data=data, ax=ax[1])
	plt.savefig('fig9.png',dpi=100)
	fig10=sns.countplot(x='region',hue='smoker', data=data, ax=ax[2])
	plt.savefig('fig10.png',dpi=100)

12.	# Import necessary libraries

SL# Reference	Applied Code

	import numpy as np
	from sklearn.model_selection import train_test_split
	from sklearn.linear_model import LinearRegression
	from sklearn.metrics import mean_squared_error,
	mean_absolute_error,r2_score
	
	#one hot encoding
	data_encoded = pd.get_dummies(data, columns=['sex','region','smoker'])
	#features and target selection
	feature = data_encoded.drop(['charges','children'], axis = 1)
	target = data_encoded.charges

13.	# Split the data into training and testing sets
	X_train, X_test, y_train, y_test = train_test_split(feature, target,
	                                                    test_size=0.2,
	                                                    random_state=42)

14.	X_train.shape, X_test.shape

15.	model = LinearRegression()
	model.fit(X_train, y_train)

16.	# Make predictions on the testing set
	y_pred = model.predict(X_test)
	y_pred

17.	MAE = mean_absolute_error(y_test, y_pred)
	print("MAE = ", MAE)
	MSE = mean_squared_error(y_test, y_pred, squared=True)
	print("MSE = ", MSE)
	RMSE = mean_squared_error(y_test, y_pred, squared=False)
	print("RMSE = ", RMSE)
	r2 = r2_score(y_test, y_pred)#Goodness of fit-coefficent of determination
	print("r_squared = ", r2)

18.	y_pred = model.predict(X_test)
	price = pd.DataFrame({"Charges_actual":y_test,
	                   "Charges_predicted": y_pred})
	price.head(10)

19.	fig14=plt.scatter(price['Charges_actual'],price['Charges_predicted'], color = 'purple')
	fig15=plt.plot(price['Charges_actual'],price['Charges_actual'], color = 'blue')

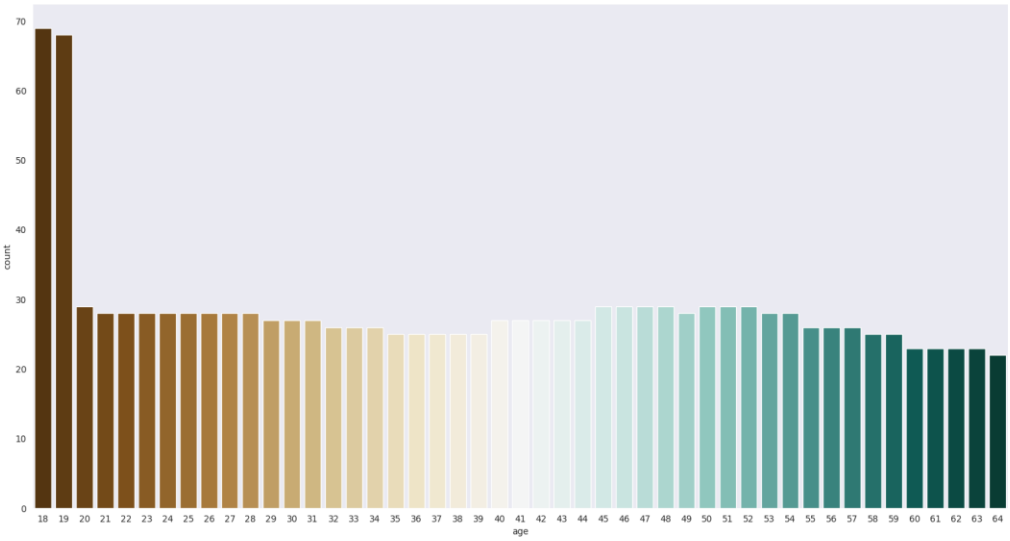
SL# Reference	Applied Code

	plt.show()
	plt.tight_layout()
	plt.savefig('fig14.png')

5. Findings

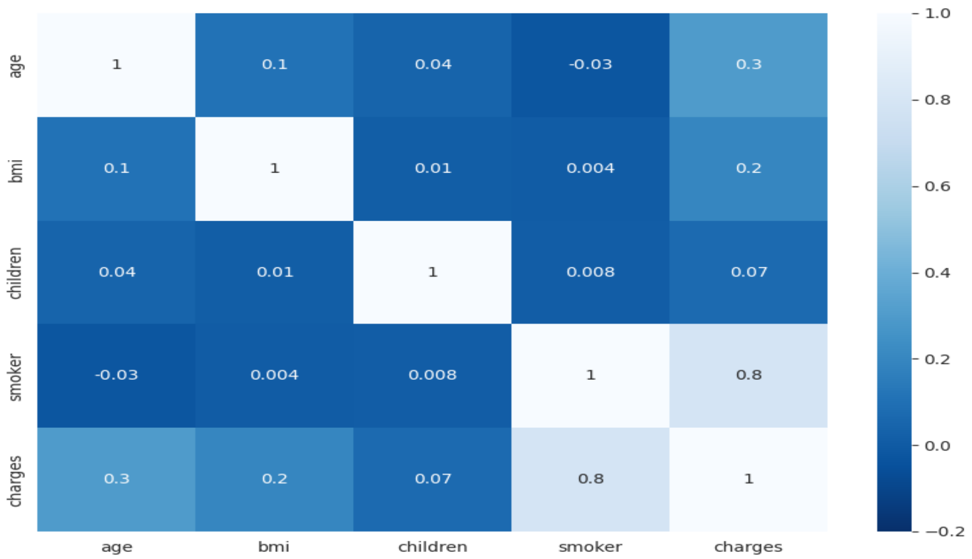
Finding of the analysis are given bellow considering the reference number from table-1.

Output-1: Reference#6



The figure in outpur-1 presented distribution of age group from the given dataset. The figure showed the age group of 18, there were 69 individuals with insurance coverage, which represents 5.2% of the total sample. The valid percent is also 5.2% since there are no missing or invalid data for this age group. The cumulative percent of insurance coverage up to and including the age group of 18 is also 5.2%. Similarly, for the age group of 19, there were 68 individuals with insurance coverage, which represents 5.1% of the total sample. Reversely, for the age group of 60, 61, 62, 63, there were 23 individuals with insurance coverage, which represents 1.7% of the total sample. Overall, the data suggests that insurance coverage is relatively consistent across age groups, with coverage ranging from 1.6% to 5.2%.

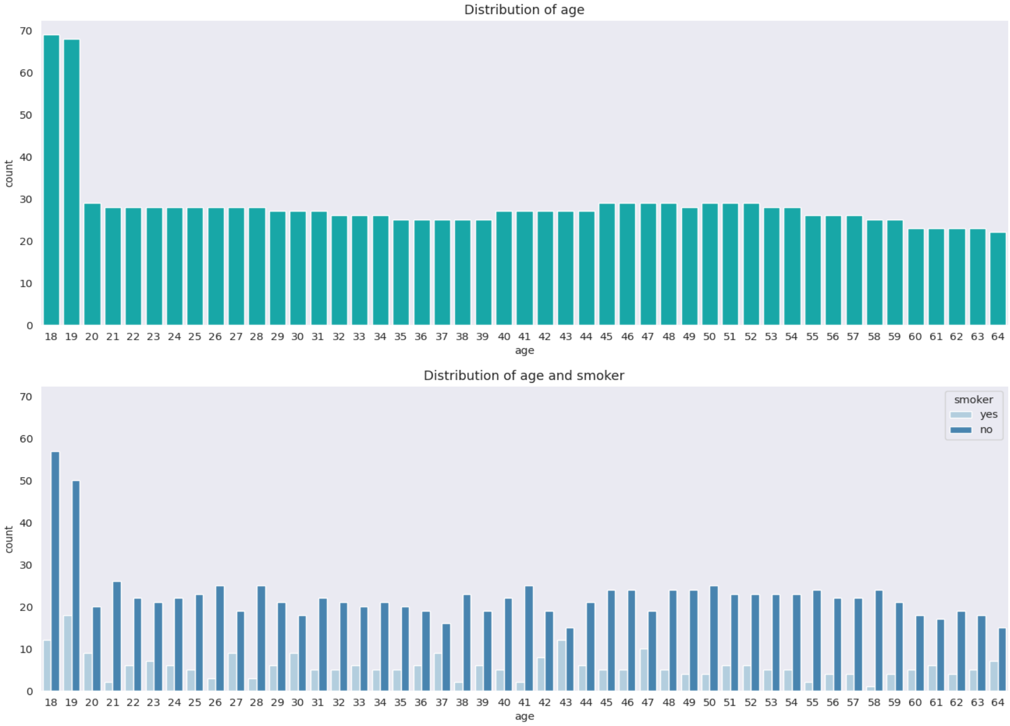
Output-2: Reference#8



The figure in output-2 showed the heat map of the Pearson correlation coefficients between the variables age, BMI, children, charges, and smoking status (smoke). Pearson correlation measures the strength and direction of the linear relationship between two variables, ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no correlation.

- Age and charges have a moderate positive correlation (r = 0.3). This means that as age increases, charges tend to increase as well, but not perfectly.
- BMI and charges have a weak positive correlation (r = 0.2). This means that there is a slight tendency for charges to increase as BMI increases, but the relationship is not very strong.
- Children and charges have a very weak positive correlation (r = 0.07). This suggests that having more children may slightly increase medical charges, but the effect is not substantial.
- **Smoke and charges have a very strong positive correlation (r = 0.8), indicating that there is a strong relationship between smoking and higher medical charges.**
- Age and BMI, as well as age and children, have weak positive correlations (r = 0.1 and r = 0.04, respectively), suggesting that there is some tendency for these variables to increase together, but the relationships are not very strong.
- All other correlations are very weak (less than 0.05) and are not practically significant.

Output-3: Reference#9



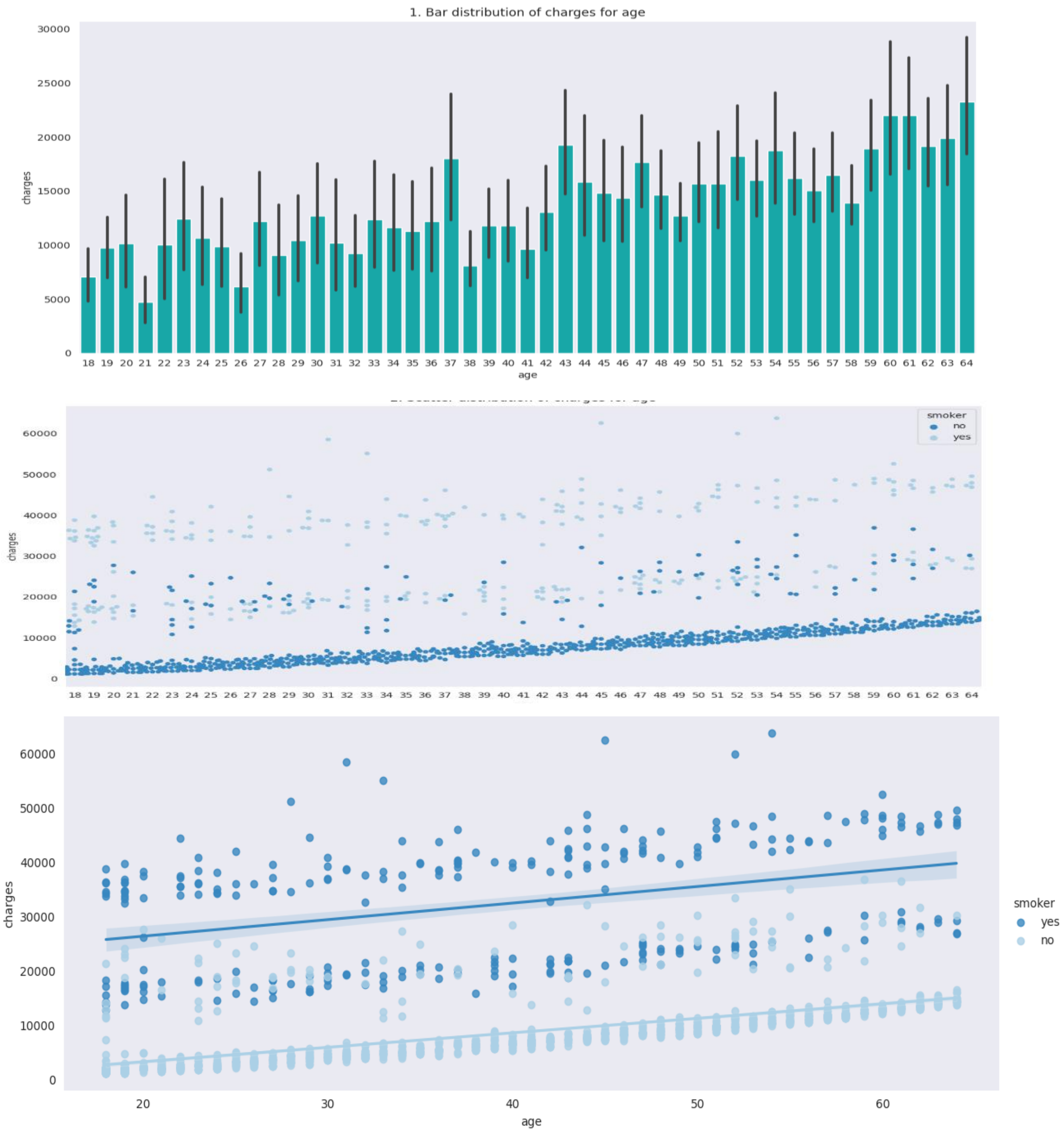
The output-3: reference#9 showed the joint distribution char of age and smoke variables that indicated the detailed analysis of the relationship between age and smoking status. Some of the key observations and insights from the charts are:

- The total count of individuals in the table is 1338, with 1064 non-smokers and 274 smokers.
- The majority of individuals in the dataset are non-smokers, with the highest number of non-smokers in the 18-19 age range.
- The number of smokers is relatively low compared to non-smokers, with the highest number of smokers in the 18-19 age range.
- The number of non-smokers decreases as age increases, with a corresponding increase in the number of smokers.
- The number of smokers peaks at age 19 and then gradually declines with increasing age.
- The lowest number of smokers is observed in the age range of 58-64.

Overall, the chart showed a clear relationship between age and smoking status, with a higher prevalence of smoking among younger individuals and a decline in smoking with increasing age.

The data could be useful for public health officials, insurance company and policymakers in developing targeted interventions to reduce smoking among younger age groups and to promote smoking cessation among older individuals.

Output-4: Reference #10

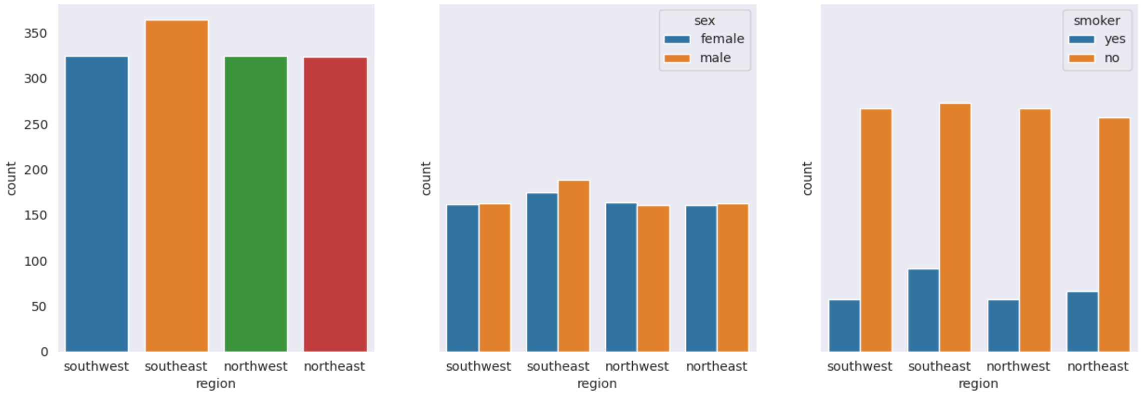


The output-4: reference#10 has produced 3 output plots of which first one is on relationship between ages and charges and next two on relationship between charges and smoking.With increasing age, charges increases. we can relate this data with real situation as we know that with increasing age, the following problems come out-our immunity to prevent disease decreases. so different diseases attacks changes our tissues, organs as they slowly lose function. For these reasons, with increasing age, people easily and more frequently become ill and different disease occurs at the same time which increase their charges.

From the scatter plot (2), we can see that, lower limit and upper limit of charges increases with the increasing age. we can vaguely see three different range of charges increasing on this plot with age starting from 2000, 15000 and 35000 charges. We can assume them low-cost, medium-cost and high-cost disease. With increasing age, severity of these diseases increases. As a result, their charges increase proportionally. Most non-smokers have low-cost disease with some outliers found in medium-cost charges. but smokers require medium to higher charges.

In the plot (3), two regression lines for smokers and non-smokers are plotted, which are straight positively growth line.

Output-5: Reference #11

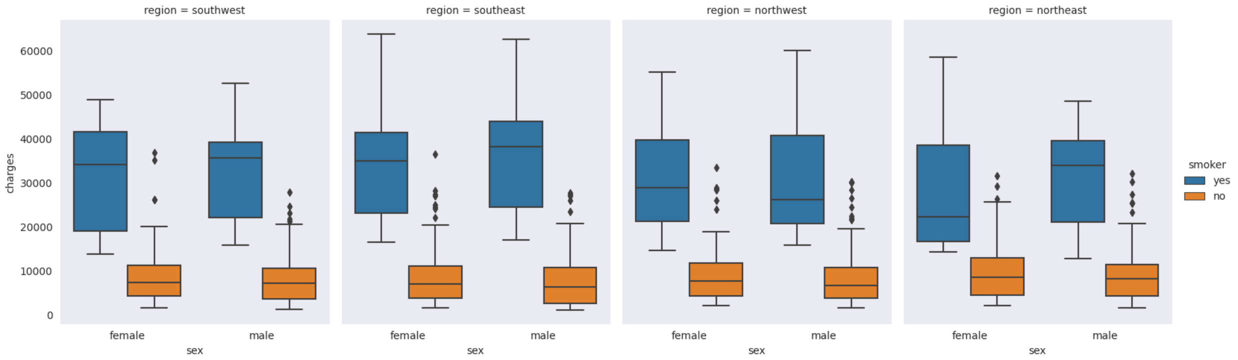


The first figure from the Output-5: Reference #11 showed the "Frequency" distribution of regional observation in the dataset. For example, "northeast" was observed 324 times, "northwest" was observed 325 times, "southeast" was observed 364 times, and "southwest" was observed 325 times. It also found that the percentage of the total observations that each value represents. For example, "northeast" represents 24.2% of the total observations, "northwest" represents 24.3% of the total observations, "southeast" represents 27.2% of the total observations, and "southwest" represents 24.3% of the total observations.

The second figure from the Output-5: Reference #11 showed a cross-relationship of the variables "region" and "sex". It shows the number of observations for each combination of values of these variables. This figure provides information about the distribution of observations across the "region" and "sex" variables. It shows that the number of observations is similar for males and females, with 676 male observations and 662 female observations. It also shows that the number of observations is similar across the four regions, with each region having approximately 25% of the total observations. Finally, it provides information about the distribution of observations across combinations of "region" and "sex", which may be useful for analyzing relationships between these variables.

The third figure from the Output-5: Reference #11 showed a cross-tabulation of the variables "region" and "smoker". It shows the number of observations for each combination of values of these variables. This figure provides information about the distribution of observations across the "region" and "smoker" variables. It shows that the majority of observations were non-smokers, with 1064 non-smoker observations and 274 smoker observations. It also shows that the number of observations is similar across the four regions, with each region having approximately 25% of the total observations. Finally, it provides information about the distribution of observations across combinations of "region" and "smoker", which may be useful for analyzing relationships between these variables. For example, the table shows that the southeast region had the highest number of smokers with 91 observations, while the northeast region had the lowest number of smokers with 67 observations.

Output-6: Reference #12



Output-6: Reference #12 showed the plot to see the charges for men and women from different region and found that no. of patients from Southeast region is highest among other regions. So male and female patients of this region is higher than other regions. So, the charges for the smoker patients are higher than other region's patients, where non-smokers charges data for this region are comparable to other regions data. Actually, from correlation methods, we have concluded that region and charges are negatively correlated.

Output-7: Reference #17

SL#	Name of model	Value
1	MAE	4222.908401655443
2	MSE	34142364.80180487
3	RMSE	5843.146823570743
4	r_squared	0.780079589226054

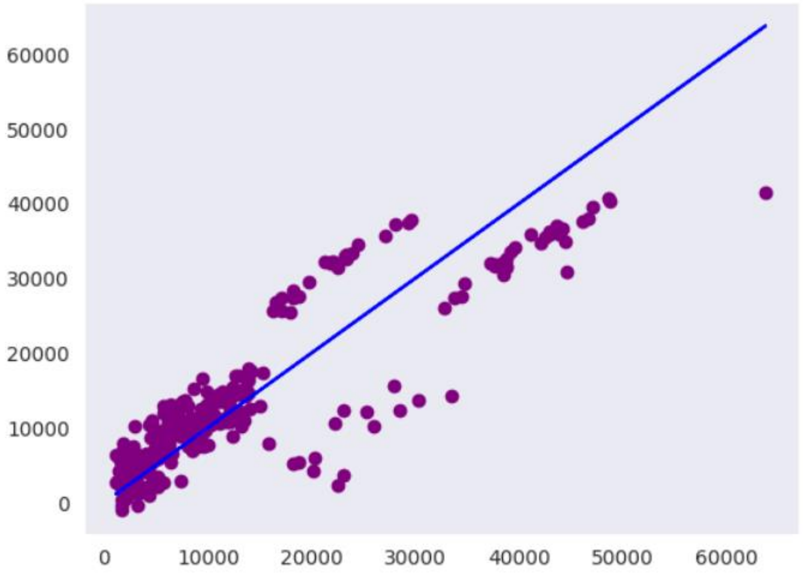
Output-7: Reference #17 calculate the MAE, MSE, RMSE and r_squared and we found:

- The MAE (Mean Absolute Error) value of 4222.91 indicates that the average difference between the actual insurance prices and the predicted insurance prices by the model is $4222.91.
- The MSE (Mean Squared Error) value of 34142364.80 is the average of the squared differences between the actual and predicted insurance prices, and indicates the overall magnitude of the errors in the predictions.
- The RMSE (Root Mean Squared Error) value of 5843.15 is the square root of the MSE, and represents the standard deviation of the differences between the actual and predicted insurance prices.
- The R-squared value of 0.780 indicates that 78% of the variability in the insurance prices can be explained by the independent variables included in the model.

Output-8: Reference #18 & 18

ID	**Charges_actual**	**Charges_predicted**
764	9095.06825	8588.336412
887	5272.17580	7548.869743
890	29330.98315	37426.741978
1293	9301.89355	8700.446936
259	33750.29180	27452.503404
1312	4536.25900	10891.857052
899	2117.33885	615.467515
752	14210.53595	17459.856040
1286	3732.62510	1527.036494
707	10264.44210	10469.434539

Output-8: Reference #18 & 19 calculated actual vs predicted prices showed in the above table. We found the predicted charges are not very accurate as there is a large difference between actual and predicted charges for some observations. For example, for the observation with Charges_actual = 890, the predicted charges are 37,426.74 which is significantly higher than the actual charges of 29,330.98. Similarly, for the observation with Charges_actual = 752, the predicted charges are 17,459.86 which is significantly higher than the actual charges of 14,210.54. If the objective is to accurately predict charges, then the model needs to be improved. This may involve identifying and adding relevant features, exploring different algorithms or hyperparameters, or using more data.



Linear between actual vs predicted charges.

6. Conclusion:

Overall, **the model seems to have a decent performance**, with a relatively low MAE and RMSE values, indicating that the model's predictions are relatively accurate. The R-squared value is also reasonably high, suggesting that the model explains a significant portion of the variability in the insurance prices. However, further analysis and evaluation may be needed to determine the model's effectiveness and suitability for its intended purpose.

It can be concluded that the assignment has assisted the students to apply machine learning techniques to analyze insurance data and build a model to predict insurance charges based on demographic and lifestyle factors. The report has summarized the findings and conclusions of the analysis and significance of the model applied. The methodology involved gathering open-source data on insurance pricing, importing it into a Jupyter notebook using Python, and performing the necessary analysis using appropriate code.