Linear Regression Modeling: Predicting Insurance Costs

In this project, I'll be delving into the Medical Cost Data Set sourced from Kaggle (https://www.kaggle.com/datasets/mirichoi0218/insurance?resource=download), which provides a comprehensive view of individual medical insurance bills, including associated demographic and personal attributes of the recipients. My primary focus will be tackling a linear regression problem, where I decipher the intricate relationship between these diverse characteristics and the total medical cost. With this cost being a continuous and positive numerical value, it aligns perfectly with the choice of employing linear regression. The mission of this project is to construct an optimal predictive model that can estimate medical expenses based on patient information. The significance of this project lies in its potential to aid hospitals in revenue projection and the strategic planning of essential healthcare procedures for their patient population.

In [127…
```
import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.model_selection import train_test_split
```

In [128…
```
%cd ~/Desktop
insurance = pd.read_excel("insurance.xlsx", header=0, skiprows=1)
```

```
/Users/palwinderdhillon/Desktop
```

Exploring the Data

Now, I will upload the explore the data and check for missing values. Below is documentation for the features in the dataset:

- `age` : age of primary beneficiary
- `sex` : insurance contractor gender, female, male
- `bmi` : Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
- `children` : Number of children covered by health insurance / Number of dependents
- `smoker` : Smoking
- `region` : the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- `charges` : Individual medical costs billed by health insurance

In [129…
```
insurance.head()
```

Out[129]:

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

In [130…
```
insurance.describe()
```

Out[130]:

	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

Here's what I can interpret from the statistics for the numerical columns:

- Age (age): The age of individuals in the dataset ranges from 18 to 64 years, with a mean age of approximately 39.2 years.
- BMI (bmi): The body mass index (BMI) ranges from approximately 15.96 to 53.13, with a mean BMI of approximately 30.66.
- Children (children): The number of children/dependents ranges from 0 to 5, with an average of approximately 1.09.
- Charges (charges): The medical charges range from 1121.87 to 63770.43, with a mean charge of approximately 13270.42.

In [131…
```
insurance.dtypes
```

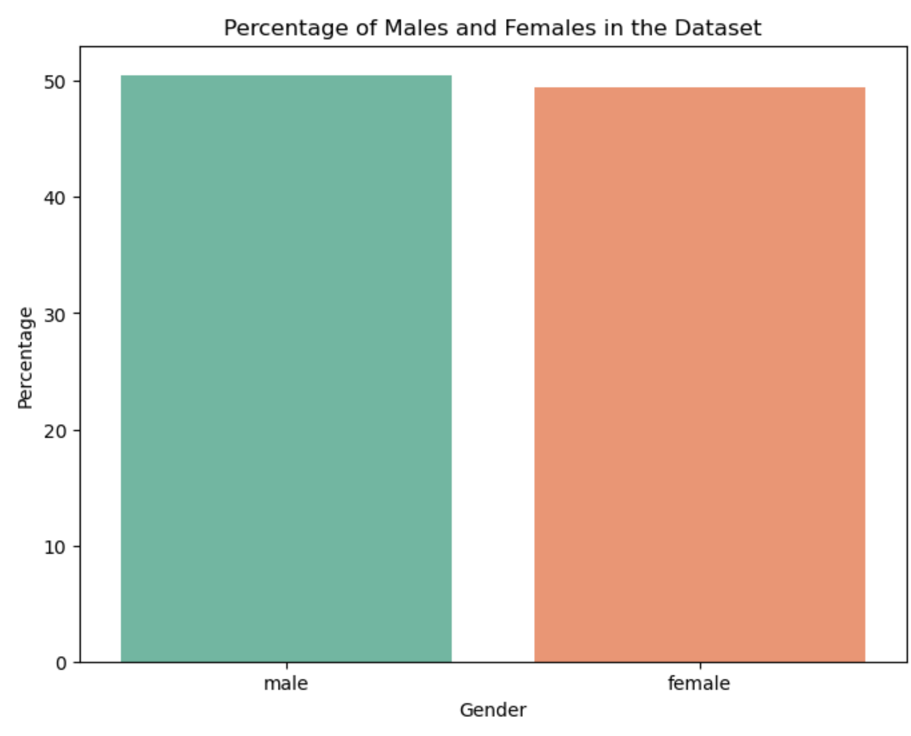
Out[131]:
```
age           int64
sex          object
bmi         float64
children      int64
smoker       object
region       object
charges     float64
dtype: object
```

In [132…
```
insurance.isnull().sum()
```

Out[132]:
```
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

I can see there are not any missing values in the dataset.

In [133…
```
import matplotlib.pyplot as plt
import seaborn as sns

gender_counts = insurance['sex'].value_counts()

gender_percentages = (gender_counts / gender_counts.sum()) * 100

plt.figure(figsize=(8, 6))
sns.barplot(x=gender_percentages.index, y=gender_percentages.values, palette="Set2")
plt.title('Percentage of Males and Females in the Dataset')
plt.xlabel('Gender')
plt.ylabel('Percentage')
plt.show()
```



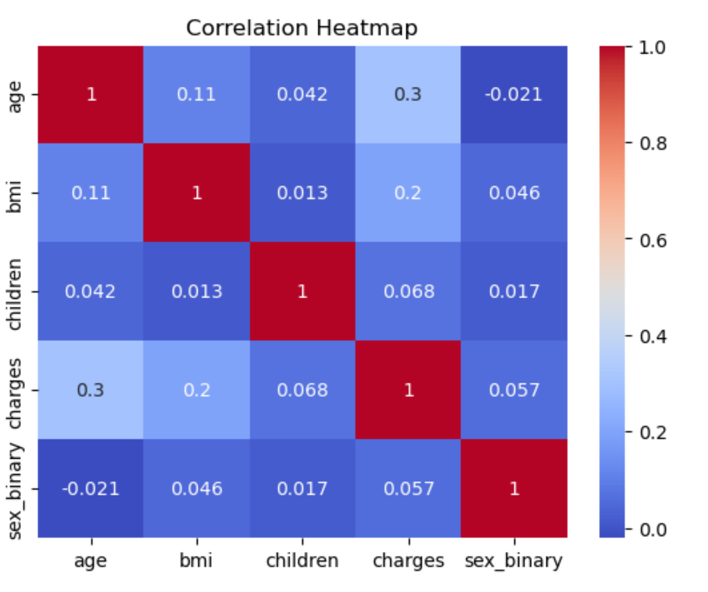
The gender distribution in the dataset is nearly equivalent.

In [134…
```
insurance['sex_binary'] = insurance['sex'].map({'female': 0, 'male': 1})

correlation_coefficient, p_value = stats.pointbiserialr(insurance['sex_binary'], insurance['charges'])

print(f"Point-Biserial Correlation Coefficient: {correlation_coefficient:.2f}")
print(f"P-value: {p_value:.4f}")
```

```
Point-Biserial Correlation Coefficient: 0.06
P-value: 0.0361
```

The value of 0.06 is quite close to zero, which suggests that there is very little linear relationship between gender (sex) and medical charges in the dataset.

In [135…
```
correlation_matrix = insurance.corr()


sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()
```



- As expected, I can see a moderate relationship of .3 where when the age of individuals in the dataset increases, the medical charges tend to increase as well.
- bmi is also moderately related with charges with a correlation coefficient of .2.

In [136…
```
import scipy.stats as stats

correlation, p_value = stats.pointbiserialr(insurance['smoker'].map({'yes': 1, 'no': 0}), insurance['charges'])

print(f"Point-Biserial Correlation: {correlation:.2f}")
```

```
Point-Biserial Correlation: 0.79
```

The positive sign of the correlation coefficient (0.79) indicates a strong positive linear relationship between being a smoker and medical charges. In other words, individuals who are smokers tend to have significantly higher medical charges compared to non-smokers in the dataset.

Based on the correlations I have investigated in this section of the project, I have chosen to include the following predictors in my linear regression that will predict `charges` :

- `age`
- `bmi`
- `smoker`

I chose these three predictors due to their positive and strong correlations to higher costs.

Dividing the Data

Now, I will divide the insurance dataset into:

1. A training set that will be used to estimate the regression coffecients
2. A test set that will be used to assess the predictive ability of the model

I will be transforming y (`charges`) using log to make extreme values less pronounced.

In [137…
```
y_log = np.log(y)

insurance['is_smoker'] = (insurance['smoker'] == 'yes').astype(int)

X = insurance[["age", "bmi", "is_smoker"]]
y = insurance["charges"]

X_train, X_test, y_train, y_test = train_test_split(X, y_log, test_size=0.25, random_state=1)
```

Building the Model

In [138…
```
model = LinearRegression()
model.fit(X_train, y_train)
```

Out[138]:
```
LinearRegression()
```

In [139…
```
intercept = model.intercept_
print(intercept)
```

```
7.135077798791309
```

The intercept of approximately 7.1351 in this linear regression model represents the estimated value of the log-transformed medical charges (the dependent variable) when all the predictor variables (age, BMI, and smoker status) are set to zero. This intercept does not carry meaning because BMI can not be zero.

In [140…
```
slope = model.coef_
print(slope)
```

```
[0.03391475 0.01056129 1.54615728]
```

Here are the interpretations of the slopes of the linear regression:

1. Age (0.03391475): For each one-year increase in age, the predicted log-transformed medical charges are expected to increase by approximately 0.0339, holding all other factors constant. This coefficient indicates that older individuals tend to have higher log-transformed medical charges.

2. BMI (0.01056129): For each one-unit increase in BMI (Body Mass Index), the predicted log-transformed medical charges are expected to increase by approximately 0.0106, holding all other factors constant. This coefficient suggests that higher BMI is associated with slightly higher log-transformed medical charges.

3. Smoker Status (1.54615728): Being a smoker (compared to being a non-smoker) is associated with a substantial increase in log-transformed medical charges. On average, smokers are predicted to have log-transformed medical charges approximately 1.5462 units higher than non-smokers, holding all other factors constant. This is a significant positive effect, indicating that smoker status has a strong impact on medical charges.

Residual Diagnostics

Now, I will examine the residuals to evaluate the linear regression model.

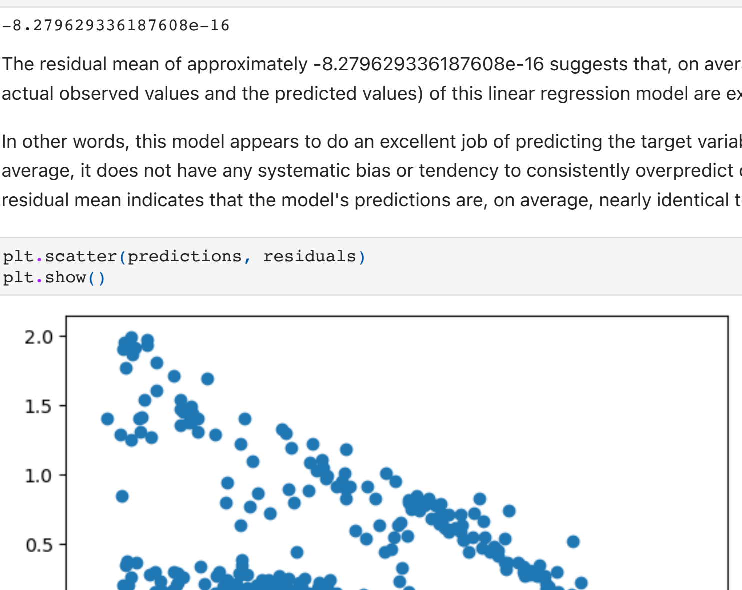
In [141…
```
model.fit(X_train, y_train)
predictions = model.predict(X_train)
residuals = y_train - predictions
residual_mean = residuals.mean()
print(residual_mean)
```

```
-8.279629336187608e-16
```

The residual mean of approximately -8.279629336187608e-16 suggests that, on average, the residuals (the differences between the actual observed values and the predicted values) of this linear regression model are extremely close to zero.

In other words, this model appears to do an excellent job of predicting the target variable (likely "charges" in this case) since, on average, it does not have any systematic bias or tendency to consistently overpredict or underpredict the actual values. The small residual mean indicates that the model's predictions are, on average, nearly identical to the actual observed values.

In [142…
```
plt.scatter(predictions, residuals)
plt.show()
```



The residuals suggest some violations to the assumptions of linear regression. As fitted values get larger, the residuals trend away from zero and go downwards. I would expect an even band, centered around zero. This does not necessarily make the model predictions unusable, but it puts into question the linear regression assumptions.

Mean Squared Errors & Final Model Evaluation on Test data: I will now evaluate the final model by seeing how it performs on test data, which will represent a set of new observations that it hasnt been trained on.

In [143…
```
train_pred = model.predict(X_train)
test_pred = model.predict(X_test)

train_mse = mean_squared_error(y_train_log, train_pred)
test_mse = mean_squared_error(y_test, test_pred)

print(train_mse)
print(test_mse)
```

```
0.21844590655300025
0.20925414547131072
```

Here are the interpretations of the mean squared errors:

- The training MSE of approximately 0.2184 suggests that, on average, the model's predictions on the training data are off by about 0.2184 units (in terms of squared log-transformed charges). This represents the model's performance on the data it was trained on.
- The test MSE of approximately 0.2093 suggests that, on average, the model's predictions on the test data are off by about 0.2093 units (in terms of squared log-transformed charges). This represents the model's performance on new, unseen data.
- The goal in model evaluation is to have lower MSE values. In this case, the test MSE is slightly lower than the training MSE, which can be a positive sign. It indicates that the model is not significantly overfitting the training data, as the test performance is similar.

Coefficient of Determination:

In [144…
```
R2 = r2_score(y_train, predictions)
print(R2)
```

```
0.7421118855283422
```

- The R2 value of 0.7421 indicates that the linear regression model explains approximately 74.21% of the variance in the medical charges. In other words, about 74.21% of the variability in the charges can be accounted for by the predictor variables (age, BMI, and smoker status) included in the model.

- R2 values range from 0 to 1, where 0 means that the model does not explain any variance, and 1 means that the model explains all the variance. In this case, an R2 value of 0.7421 suggests that the model captures a substantial portion of the variation in medical charges.

Conclusion

In this linear regression analysis, I aimed to develop a predictive model for medical charges based on a dataset that includes patient age, BMI, and smoker status. After preprocessing the data, including log-transforming the target variable for improved modeling, I built a linear regression model. The model yielded insightful coefficients: age and BMI showed relatively small but positive effects on medical charges, while being a smoker had a significant positive impact. The intercept indicated the estimated log-transformed charges when all predictors were at their baseline values, although its interpretation does not carry significance.

The model performed well in terms of explaining the variance in medical charges, with an R-squared (R2) value of approximately 0.7421, signifying that it accounts for about 74.21% of the variance in charges. Additionally, the mean squared error (MSE) values of 0.2184 on the training dataset and 0.2093 on the test dataset indicated that the model's predictions were generally close to the actual charges. The small residual mean of nearly zero further emphasized that the model's predictions align well with the observed values.

In conclusion, this linear regression model offers valuable insights into the factors influencing medical charges. Smoker status emerged as a particularly significant determinant, while age and BMI played smaller roles. The model's relatively high R2 and low MSE values suggest that it provides a reasonable basis for predicting medical charges based on these factors.

Next steps: Further refinement and validation may be necessary, and consideration of other relevant variables could enhance predictive accuracy for real-world applications.