Objective

In this document, I intend to examine the data of a health insurance firm to anticipate future insurance claims and detect factors that contribute to higher medical expenses. The insurance database comprises 1,338 policyholders with seven features that describe their demographic and health-related information, such as age, gender, BMI, and smoking habits. The target variable that we aim to forecast is the total claim amount or the fees that the insurance company was charged. To create a more precise cost projection, I will construct a model for each of the four regions where the policyholders of the company reside - the Southwest, Southeast, Northwest, and Northeast regions.

Data Overview

Data Overview:

**Code:**

data = pd.read\_csv("/Users/Jorden/PycharmProjects/Insuranceproject/insurance.csv")  
print(f"There a {data.shape[**0**]} observations and {data.shape[**1**]} columns in the data set")  
print(f"There are {data.isna().sum().sum()} missing values from the data")

**Output:**

A picture containing text, orange, dark

Description automatically generated

Summary Statistics

**Code:**

#Numerical Data-types  
pd.set\_option('display.max\_columns'**,** None)  
print(data.describe().T)

Graphical user interface, text, application, chat or text message

Description automatically generated**Output:**

Text

Description automatically generated

Summary of Catagorical Columns

**Code:**

#Object Data-types  
#Smoker data is yes/no format, want to show: Smoker/Non-Smoker  
data['Smoker'] = data['Smoker'].apply(lambda x: 'Smoker' if x == 'yes' else 'Non-Smoker')  
print(data.select\_dtypes(include='object').describe().T)

A screenshot of a computer

Description automatically generated with medium confidence**Output:**

**Code:**

**Code:**

#Of the object Data-types, find the avg charge(insurance payout) for each of Sex, Smoker/Non, Region  
catagories\_lst = data.select\_dtypes(include=['object']).columns.tolist()  
print(catagories\_lst)  
for column in catagories\_lst:  
 observation = data[column].value\_counts()  
 avg\_claim = data.groupby(column)['Charges'].mean()  
 title1 = "Number of Policy Holders"  
 title2 = "Average Claim Amount"  
 display(pd.DataFrame({title1: observation**,** title2: avg\_claim})\  
 .sort\_values(title1**,** ascending=False))

**Output:**

**A picture containing text

Description automatically generated**

**Text

Description automatically generated with medium confidence**

**A screenshot of a computer

Description automatically generated with medium confidence**

Exploratory Data-Analysis

**Code:**

#Insurance Claim Amount Based on Region  
sns.boxplot(x='Region'**,** y='Charges'**,** data=data)\  
 .set(title="Insurance Claim Amount Based on Region")

**Output:**

**Chart, box and whisker chart

Description automatically generated**

Comments

**The boxplots reveal that insurance charges in all regions have a positive skewness with a few significant outliers. The Southeast region has the most variability in insurance charges, and it includes the highest claim in the data set, amounting to more than $63,000, and the lowest claim at $1,121. On the other hand, the Northeast region has the highest median cost overall, but as the notches in the boxplots overlap, the median claim amounts are likely not considerably distinct.**

**Code:**

#Insurance Claim Amount Based on Smoking Status  
sns.boxplot(x='Smoker'**,** y='Charges'**,** data=data)\  
 .set(title="Insurance Claim Amount Based on Smoking Status")

**Output:**

**Chart, box and whisker chart

Description automatically generated**

**Code:**

#Probabilty of Different Claim Size Depending on Smoking Status  
sns.displot(data**,** x='Charges'**,** hue='Smoker'**,** stat='probability')\  
 .set(title='Probabilty of Different Claim Size Depending on Smoking Status')

**Output:**

**Chart, histogram

Description automatically generated**

Comments

**The median claim amount for non-smokers is considerably lower at around $7,300 when compared to smokers, whose median claim is over $34,000. Additionally, the distribution of insurance costs between these two groups displays noteworthy differences. The distribution of insurance costs for non-smokers is positively skewed, whereas the distribution for smokers has a wider spread in the boxplot and a bimodal distribution with two peaks near $20,000 and $40,000, indicating more variability in claim amounts.**

**Code:**

#Split into age groups, find avg claim amount, plot based on smoking-status/age  
group\_df = data.copy()  
group\_df["Age\_Group"]=['18 to 29 years' if i<**30** else '30 to 44 years' if (i>=**30**)&(i<**45**) else  
 '45 to 59 years' if (i>=**45**)&(i<**60**) else '60 and over' for i in data['Age']]  
group\_df = group\_df.groupby(['Age\_Group'**,**'Smoker'])['Charges'].mean()  
group\_df = group\_df.rename('Charges').reset\_index().sort\_values('Smoker'**,** ascending=True)  
sns.barplot(group\_df**,** x='Age\_Group'**,** y='Charges'**,** hue='Smoker')\  
 .set(title='Average Claim Cost by Age Group and Smoking Status')

**Output:**

**Chart, bar chart

Description automatically generated**

**Code:**

sns.lineplot(data**,** x='Age'**,** y='Charges'**,** hue='Smoker')\  
 .set(title='Claim Cost by Age Group and Smoking Status')

**Output:**

**Chart

Description automatically generated**

Comments

**On average, insurance charges for smokers are significantly higher than non-smokers across all age groups. Additionally, as age increases, we observe higher claim amounts, with the most expensive claims being in the age group of 60 and over.**

**Code:**

sns.scatterplot(data**,** x='BMI'**,** y='Charges'**,** hue='Smoker')\  
 .set(title='Claim Cost relation to BMI Seperated by SMoking Status')

**Output:**

**Chart, scatter chart

Description automatically generated**

Comments

**Apart from age, there is a positive correlation between claim amounts and body mass index (BMI). Specifically, among smokers, insurance charges tend to rise with increasing BMI.**

**Code:**

data['Region'] = data['Region'].apply(lambda x: 'NE' if x == 'northeast' else \  
 ('NW' if x=='northwest' else ('SE' if x=='southeast' else 'SW')))  
  
group\_plot = data.groupby(['Region'**,** 'Sex'**,** 'Smoker'])['Charges'].mean()  
#print(group\_plot)  
  
group\_plot = group\_plot.rename('Charges').reset\_index()  
  
g = sns.FacetGrid(group\_plot**,** col='Region'**,** row='Smoker'**,** hue='Smoker'**,**legend\_out=True)  
 #.set\_titles(row\_template=)  
g.map\_dataframe(sns.barplot**,** y='Charges')  
p.show()

**Output:**

**A picture containing box and whisker chart

Description automatically generated**

Comments

On average, non-smokers have smaller insurance claims than smokers, with a mean claim amount of just under $10,000 in each region, while smokers' claims vary from an average of $28,000 to over $36,000. Furthermore, the graph illustrates that among non-smokers, women have marginally higher average costs than men. Conversely, among smokers, men tend to have slightly higher charges on average than women.