**Data Analytics for Insurance Cost Data Set using Python**

In this project, you will serve as a Data Analyst and have to perform analytics operations on an insurance database that uses the below mentioned parameters.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | **Parameter** | **Description** | **Content type** | | --- | --- | --- | | age | Age in years | integer | | gender | Male or Female | integer (1 or 2) | | bmi | Body mass index | float | | no\_of\_children | Number of children | integer | | smoker | Whether smoker or not | integer (0 or 1) | | region | Which US region - NW, NE, SW, SE | integer (1,2,3 or 4 respectively) | | charges | Annual Insurance charges in USD | float | |

## **Objectives**

In this project, as a data analyst I have performed the below tasks:

* Load the data as a pandas dataframe
* Clean the data, taking care of the blank entries
* Run exploratory data analysis (EDA) and identify the attributes that most affect the charges
* Develop single variable and multi variable Linear Regression models for predicting the charges
* Use Ridge regression to refine the performance of Linear regression models.

**Below python libraries are used for the data analysis.**

* [pandas](https://pandas.pydata.org/?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMML0187ENSkillsNetwork31430127-2021-01-01) for managing the data.
* [numpy](https://numpy.org/?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMML0187ENSkillsNetwork31430127-2021-01-01) for mathematical operations.
* [sklearn](https://scikit-learn.org/stable/?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMML0187ENSkillsNetwork31430127-2021-01-01) for machine learning and machine-learning-pipeline related functions.
* [seaborn](https://seaborn.pydata.org/?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMML0187ENSkillsNetwork31430127-2021-01-01) for visualizing the data.
* [matplotlib](https://matplotlib.org/?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMML0187ENSkillsNetwork31430127-2021-01-01) for additional plotting tools.

**LET’S BEGIN THE ANALYSIS :-**

**TASK 1 – DATA IMPORTING**

**IMPORTING THE LIBRARIES**

**import** pandas **as** pd

**import** matplotlib.pyplot **as** plt

**import** numpy **as** np

**import** seaborn **as** sns

**from** sklearn.pipeline **import** Pipeline

**from** sklearn.preprocessing **import** StandardScaler, PolynomialFeatures

**from** sklearn.linear\_model **import** LinearRegression, Ridge

**from** sklearn.metrics **import** mean\_squared\_error, r2\_score

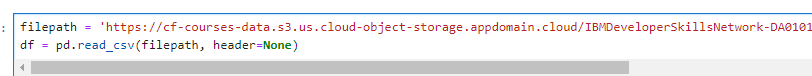
**from** sklearn.model\_selection **import** cross\_val\_score, train\_test\_split

**LOADING AND PREVIEWING THE DATA**

filepath = 'https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DA0101EN-Coursera/medical\_insurance\_dataset.csv'

**Import the data set using below command.**

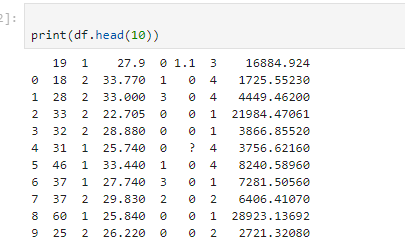
**df = pd.read\_csv(filepath, header=None)**



This command Import the dataset into a pandas dataframe. Note that there are currently no headers in the CSV file.

**Print the first 10 rows of the dataframe to confirm successful loading.**

**df.head(10)**



**DATA CLEANING**

**Add the headers to the dataframe.**

**Headers=[“age”,”gender”,”bmi”,”no\_of\_children”,”smoker”,”region”,”charges”]  
 df.columns=headers**



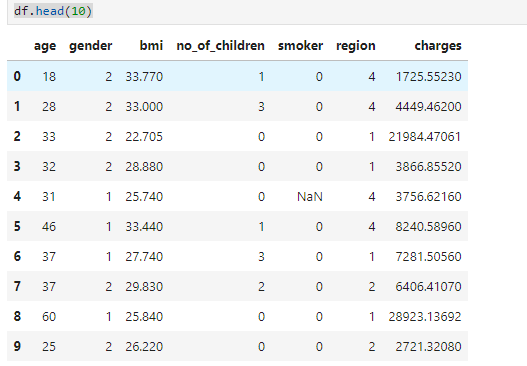
**TASK 2 : DATA WRANGLING**

**We have incorrect values as ?, so we will replace the '?' entries with 'NaN' values.**

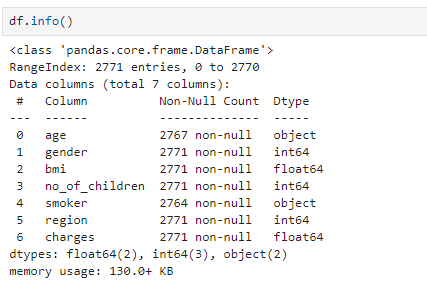
**df.replace(“?”,np.nan, inplace = True)**



**Displaying the data to recheck**

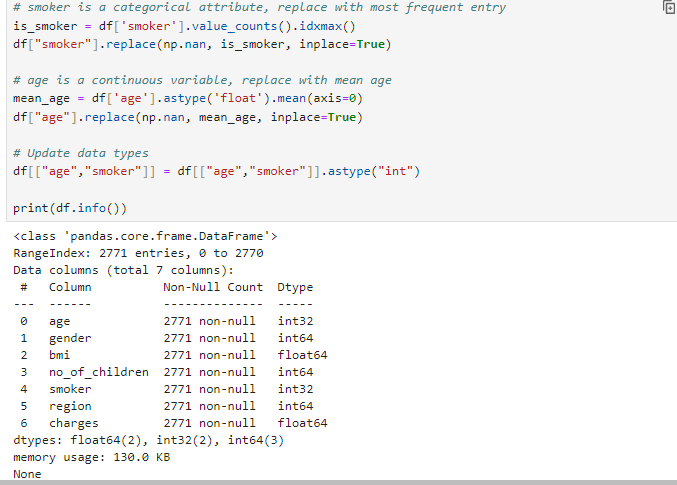


**Used dataframe.info() to identify the columns that have some 'Null' (or NaN) information.**

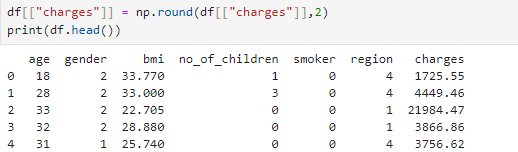


**Now, we will Handle missing data**:

* For continuous attributes (e.g., age), replace missing values with the mean.
* For categorical attributes (e.g., smoker), replace missing values with the most frequent value.
* Update the data types of the respective columns.
* Verify the update using df.info().



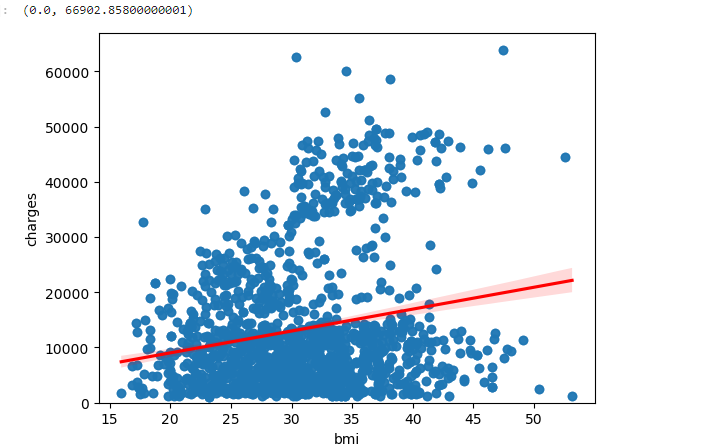
**charges**column has values which are more than 2 decimal places long. Update the charges column such that all values are rounded to nearest 2 decimal places. Verify conversion by printing the first 5 values of the updated dataframe.



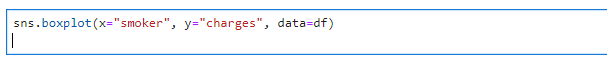
**Task 3 : Exploratory Data Analysis (EDA)**

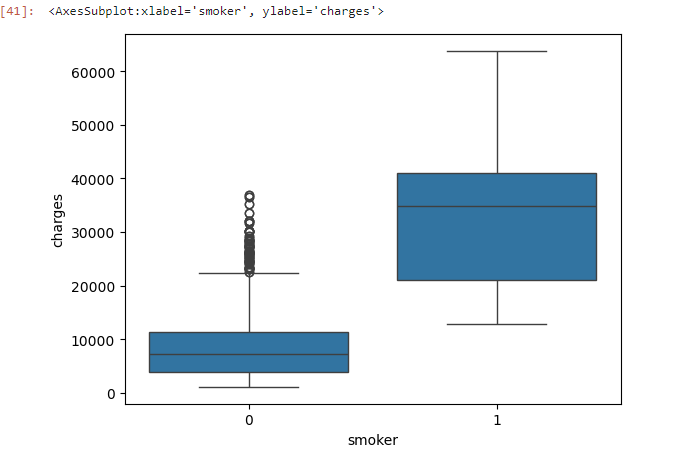
**Implement the regression plot for charges with respect to bmi**





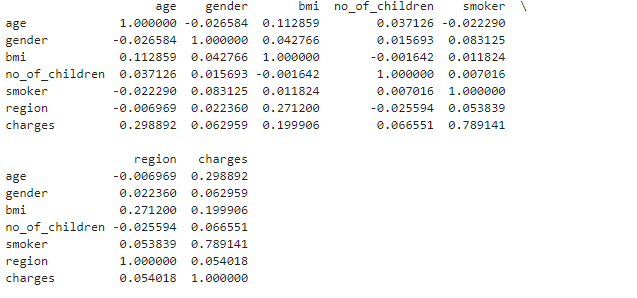
**Implement the box plot for charges with respect to smoker.**





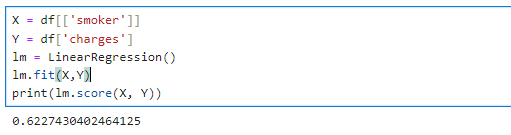
**Print the correlation matrix for the dataset.**



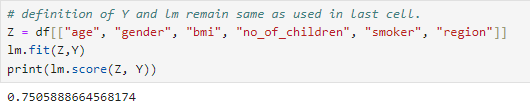


**Task 4 : Model Development**

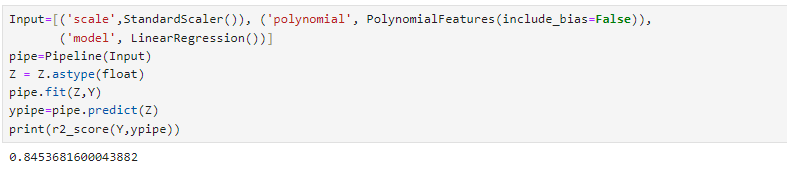
**Fit a linear regression model that may be used to predict the charges value, just by using the smoker attribute of the dataset. Print the R^2 score of this model.**



**Fit a linear regression model that may be used to predict the charges value, just by using all other attributes of the dataset. Print the R^2 score of this model. You should see an improvement in the performance.**



**Create a training pipeline that uses StandardScaler(), PolynomialFeatures() and LinearRegression() to create a model that can predict the charges value using all the other attributes of the dataset. There should be even further improvement in the performance.**

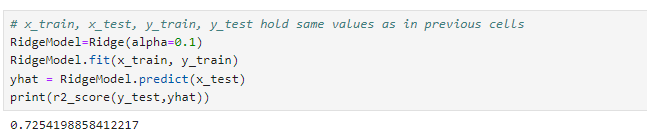


# Task 5 : Model Refinement

**Split the data into training and testing subsets, assuming that 20% of the data will be reserved for testing.**



**Initialize a Ridge regressor that used hyperparameter alpha=0.1. Fit the model using training data data subset. Print the R2 score for the testing data.**



**Apply polynomial transformation to the training parameters with degree=2. Use this transformed feature set to fit the same regression model, as above, using the training subset. Print the R2 score for the testing subset.**

