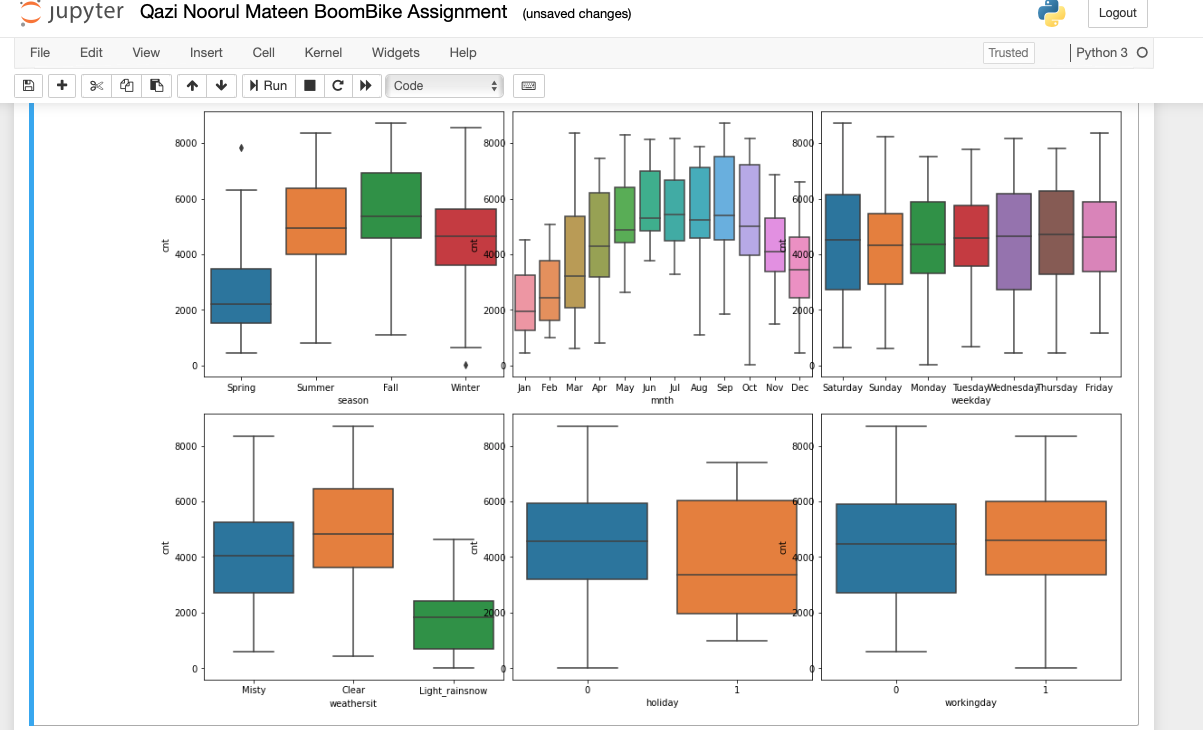
**Assignment-based Subjective Questions**

**1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?** (3 marks)



**Analysis is-**

1. Spring season has the least demand for bike.  
2. The middle months of the year (May to September) have the most demand for bike.  
3. All days of weak don’t show much variation, which implies the necessity of bike every day.  
4. In the case of rain-snow the demand for the bike is less, which is quite logical too.  
5. The demand for bike is less in holidays if we compare holidays vs non holidays.

**2.Why is it important to use drop\_first=True during dummy variable creation? (2 mark)**   
Ans- drop\_first=True helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. Number of dummies = k-1. For instance,three values furnished, unfurnished and semi furnished then we can drop one and can get result. If the data don’t lie in these two then automatically its third one (the dropped one).

**3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

Ans- Temp and atemp

**4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

Ans - By analysing the F statistics - model fit statistically significant or not? R square value - variance explained by these variables   
P value- significance of coefficient.

**5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

Ans- Temp , Light rain snow and year

**General Subjective Questions**

**1. Explain the linear regression algorithm in detail. (4 marks)**  
Ans- Linear Regression is one of the most popular and easy to apply machine learning algorithm based on supervised learning. Predictive analysis is done using this algorithm. As the name suggest linear regression algorithm shows a linear relationship between a dependent and one or more independent variable.  
The linear regression equation for this is-  
y = B̆ø + B̆1 x1 + B̆2 x2…. + B̆n xn

To apply this algorithm in python –   
Two popular packages can be used viz. statsmodel and sklearn  
  
The steps involved in the algorithm are-   
1. Create x and y  
2. Create train and test sets  
3. Train model on training set using statmodel or sklearn  
4. Evaluate the model

**2. Explain the Anscombe’s quartet in detail. (3 marks)**

Ans- This method was given buy Anscombe in 1973 to tell that graphing the data can help us to analyse the data more significantly. Apart from statistics things like mean,variance these graph supplement the analysis. As the name suggests (quartet), it comprises four datasets with nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. So he explained graphing the data is very important to conclude.

**3.What is Pearson’s R? (3 marks)**  
Ans - It is a is a numerical summary of the strength of the linear association between the variables where correlation coefficient varies between -1 and +1 where:   
r = 1 means the data is perfectly linear with a positive slope ( i.e., both variables tend to change in the same direction)  
r = -1 means the data is perfectly linear with a negative slope ( i.e., both variables tend to change in different directions)

r = 0 means there is no linear association  
r > 0 < 5 means there is a weak association  
r > 5 < 8 means there is a moderate association r > 8 means there is a strong association

**4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**Ans- Scaling is making the data to a convenient scale, like sometimes data is too much spread in magnitude so we can scale it a number between 0-100 or so. Scaling help the user to interpret it easily. Scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc*.*

**Normalization vs Standardization**

Normalization Scaling – It is also known as Min-Max scaling, use to bring all the data in range of 0 to 1. In our Linear regression machine learning model, Sklearn has MinMaxscaler function. This function can be called by - sklearn.preprocessing.MinMaxScale.   
  
However, in Standardization Scaling the values are replaced by their Z scores. It brings all the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

**5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**  
Ans- VIF (Variance Inflation Factor) is a method to judge multicollinearity among the predictor variables. An infinite VIF value indicates that the corresponding variable is expressed exactly by a linear combination of other variables i.e., perfect correlation. The value of VIF greater than 1 shows the correlation among predictors and with increasing VIF, multicollinearity increases. Generally we set the cut off for VIF as 4.   
  
  
**6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks**)   
  
Ans - Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions. It is used to check the following scenarios:  
If two data sets —  
i. come from populations with a common distribution

ii. have common location and scale  
iii. have similar distributional shapes  
iv. have similar tail behaviour