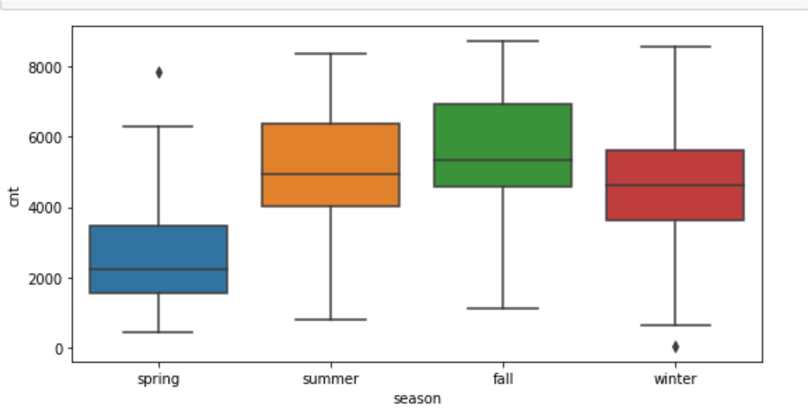
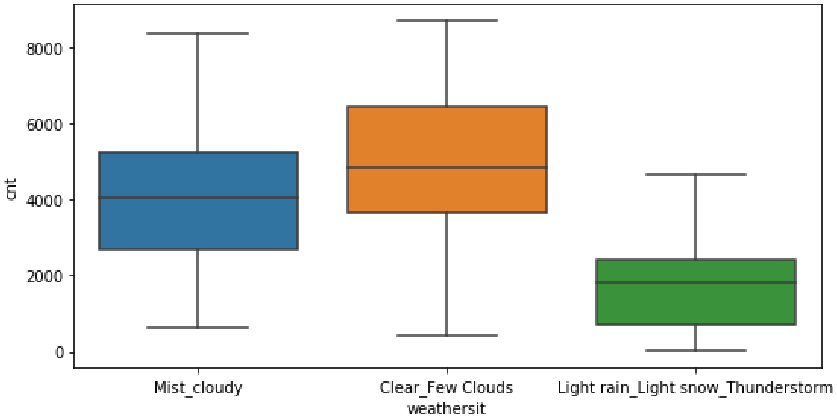
Assignment-based Subjective Questions

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

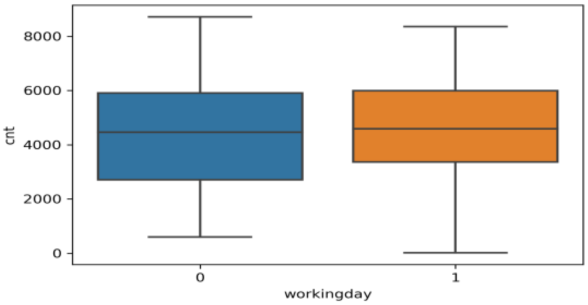
   Ans)



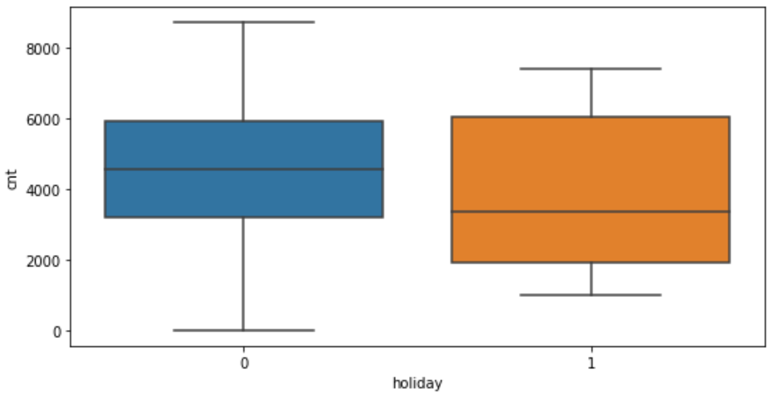
1.a)Rental bikes is highest in fall followed by summer , winter and spring



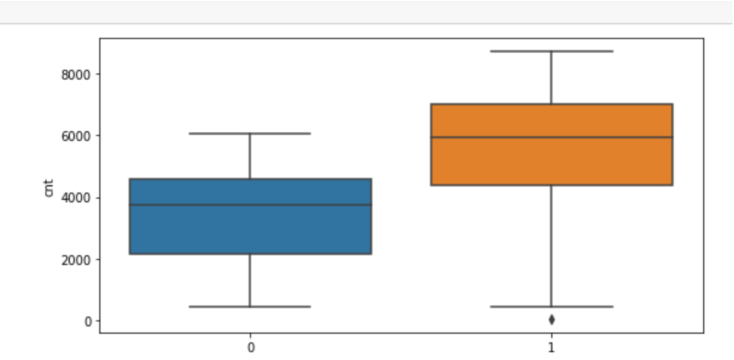
1.b) CNT is higher in clear few cluouds followed by cloudy and bad weather



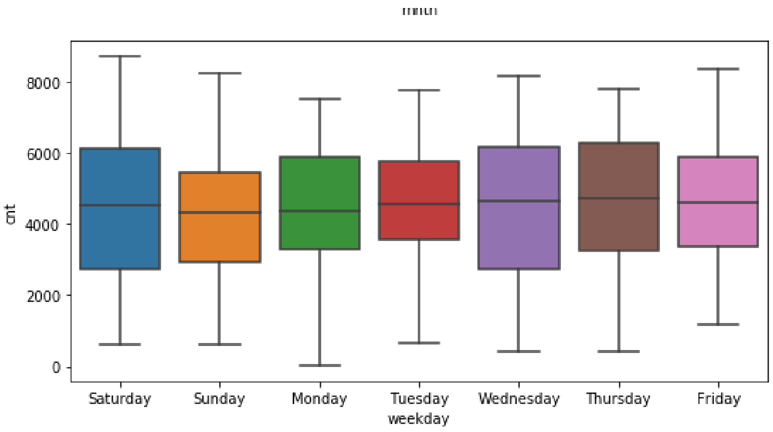
1.c) Working days have slightly higher demands than holidays



1.d) Holidays have lower demand as compared to working days



1.e) 2019 seen increase in rental bike as compared to 2018



1.f) In starting and ending of week rental bike increases and again deacreases in mid of week.

2. Why is it important to use **drop_first=True** during dummy variable creation?

   Ans2) It is important to use drop_first = True because it helps in reducing extra columns. We have also used it in our variables like wethersit, weekday, season and month.

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

   Ans3) Based on pair plot temperature has the highest correlation with the target value.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

   Ans4.1)

VARIANCE INFLATION FACTOR CODE

```
 from statsmodels.stats.outliers_influence import variance_inflation_factor

vif = pd.DataFrame()
X = X_train_rfe
vif['Features'] = X.columns
vif['VIF'] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
vif['VIF'] = round(vif['VIF'], 2)
vif = vif.sort_values(by = "VIF", ascending = False)
vif
```

4.2) **P Value CODE**

```
import statsmodels.api as sm
X_train_rfe1 = sm.add_constant(X_train_rfe)
lm = sm.OLS(y_train,X_train_rfe1).fit()
print(lm.summary())
```

4.3) **Significance Value**

High P-value, High VIF(Not Recommended)

- Low P value, Low VIF (Significant)
- High-Low: - High P, low VIF (Not Recommended)
- Low p, High VIF(Significant)

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

   Weathersift , temperature and season are top feature effecting demand of shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   1. Reading the data: import pandas and other important libraries for reading the data.

   2. Analysis of data(Exploratory Data Analysis): Analysis of the correlation between different variables.

   3. Data Preparation: Converting strings to numeric values using dummy function.

   4. Splitting the Data into Training and Testing Sets:

   5. Building a linear model 6. Residual Analysis of the train data

   7. Making Predictions Using the Final Model:Extrapolating the graph.

   8. Model Evaluation: Plotting the y test and y predicted to find the significant variables.

2. Explain the Anscombe’s quartet in detail.
   Ans) Anscombe's Quartet are variables that are similar in statistics but when plotted in scatter plots they behave differently and show some peculiarities. It is a set of four dataset.

3. What is Pearson’s R?
   Ans) It is the ratio of covariance of two variables and product of their standard deviations.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   Ans) Scaling is changing the upper and lower bounds of the dataset to a single range.
   Normalization:is scaling between 0 and 1 Standardization is scaling between -1 to 1
   Scaling is performed so that multiple variables can be viewed on the same graph with different scales.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   Ans) In the case of perfect correlation , R2 = 1 which leads to 1/(1-R2) to infinite. That implies a perfect correlation between variables

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   Ans) Quantile -Quantile plots are plots between two Quantiles(ex.25% so that number of points below 25% in a plot). It compares two different correlation plots in a variable.