Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

- The demand of bikes seems to be veritably low in the month of spring when compared with other seasons.
- The demand of bikes in 2019 is more as compared to former time i.e 2018. * Month June to Sep is the period when bike demand is high. The Month Jan is the smallest demand month.
- Bike demand is seems to drop when it's a vacation compared to normal working days.
- The demand of bike is nearly constant throughout the week. There's no significant change in bike demand with working day and non working day.
- The bike demand is high when rainfall is clear and Many shadows still demand is less in case of Light snow and light downfall. We don't have any dat for Heavy Rain Ice Pallets Thunderstorm Mist, Snow Fog, so we can't decide any conclusion. May be the company isn't operating on those days or there's no demand of bike.

Q2. Why is it important to use drop_first=True during dummy variable creation?

- Drop_first = True is important to use, as it helps in reducing the redundant column created during ersatz variable creation. thus it reduces the correlations created among ersatz variables.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

- atemp and temp both have same correlation with target variable of 0.63 which is the loftiest among all numerical variables.

Q4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

1. **Linearity :**

   Linear retrogression needs the relationship between the independent and dependent variables to be direct. We've used a brace plot to check the relation of independent variables with the Deals variable

2. **Mean of residuals:**

   Residuals is the differences between the factual(true) value and the prognosticated value. One of the hypotheticals of direct retrogression is that the mean of the residuals should be 0.

3. **Homoscedasticity Check:**

   Homoscedasticity means that the residuals have equal or nearly equal friction across the retrogression line. We compass the error terms with prognosticated terms so we can check that there shouldn't be any pattern in the error terms.

4. **No proper Multicollinearity:**

   In retrogression, multicollinearity refers to the extent to which independent variables are identified. Multicollinearity affects the portions and p- values but it doesn't affect the prognostications, perfection of the prognostications, and the virtuousness- of- fit statistics. The thing is to make prognostications, and you dont want to understand the part of each independent variable you don't need to reduce multicollinearity.

5. **No Autocorrelation of residual:**

   The residuals are autocorrelated it means that the current value is dependent of the former values and that there's a definite uncertain pattern in the Y variable that shows up in the error terms. Though it's more effective in time series data. This is usual in time series data as there's a pattern of time for illustration we can assume week of the day effect which is a veritably notorious pattern seen in stock requests where people put sweats to buy stocks more towards the morning of weekends and put some sweats to vend stocks more on Mondays. There is been great study about this trial and it's still a matter of exploration as to what factual factors beget this trend.

Q5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

1. Temperature(0.552)
2. Weather condition(weathersit) Light Snow, Light Rain Rainstorm Scattered shadows, Light Rain Scattered shadows(-0.264)
3. year (0.256)

General Subjective Questions

Q1. Explain the linear regression algorithm in detail

- Linear Retrogression is a machine learning algorithm grounded on supervised literacy. It execute's a retrogression task.
- Retrogression models a target vaticination value grounded on independent variables. It's oftenly used for getting the relationship between variables and sooth saying.
- Retrogression models differ grounded on the relationship between dependent and independent variables only by considering the number of independent variables getting used.
- Linear retrogression performs the task to prognosticate a dependent variable value(y) grounded on a given independent variable(x). So, this retrogression fashion finds out a direct relationship between x(input) and y(affair). Hence, the name is LinearRegression.

Q2. **Explain the Anscombe's quartet in detail.**

- Anscombe's Quintet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics but there are some functionality in the dataset that fools the retrogression model if are in process. They've veritably different distributions and appear else when colluded on smatter plots.
- It was constructed in 1973 by statistician Francis Anscombe to illustrate the significance of conniving the graphs before assaying and model structure, and the effect of other compliances on statistical parcels.
- There are these four data set plots which have nearly same statistical compliances, which provides same statistical information that involves friction, and mean of all x, y points in all fourdatasets.
- The four datasets can be described as
  1. Dataset 1 this fits the direct retrogression model enough well.
  2. Dataset 2 this could not fit direct retrogression model on the data relatively well as the data is non-direct.
  3. Dataset 3 shows the outliers involved in the dataset which can not be handled by direct retrogression model
  4. Dataset 4 shows the outliers involved in the dataset which can not be handled by direct retrogression model

Q3. What is Pearson's R?

- Pearson R helps us to show how explosively the direct relationship between the 2 variables. It also called as Correlation Measure. Pearson R ranges from 1 to-1. In which-1 means the variables have negative direct relation. 0 indicates that there's no relation between the variables. And incipiently 1 tells us that there's a perfect positive direct relation between the variables.

Q4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

- Scaling helps in data Pre-Processing which is applied to independent variables to homogenize the data within a certain range. It also helps in adding the speed of the computations in an algorithm.

- utmost of the times collected data set contains features largely varying in bulks, units and range. If scaling isn't used also algorithm only takes magnitude in consideration and not units hence incorrect modelling happens. thus to avoid similar problems we've to do scaling to bring all the variables to the same position of magnitude.

Difference between Normalization and Standardization

S.NO.	Normalization	Standardization
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
3.	It is really affected by outliers.	It is much less affected by outliers.
4.	Scikit-Learn provides a transformer called Min-Max Scaler for Normalization.	Scikit-Learn provides a transformer called Standard Scaler for standardization.
5.	It is Also called as Scaling Normalization	It is also called as Z-Score Normalization.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

- For perfect correlation, there should be VIF = perpetuity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 = 1, which lead to 1/(1- R2) perpetuity. * To avoid similar issue we've to drop one of the variables from the dataset which generates the perfect multicollinearity.
- An horizonless VIF value indicates that the corresponding variable may be expressed exactly by a direct combination of other variables(which show an horizonless VIF as well).

Q.6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

- Q- Q Plots(Quantile- Quantile plots) are plots of two quantiles against each other. A quantile is a bit where certain values fall below that quantile. For illustration, the standard is a quantile where 50 of the data fall below that point and 50 taradiddle above it.
- The purpose of Q- Q plots is to admit if two sets of data come from the same distribution. A 45º angle is colluded on the Q- Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.
- If the two distributions being discerned are analogous, the points in the Q – Q plot will roughly lie on the line y = x.
- If the distributions are linearly related, the points in the Q – Q plot will roughly lie on a line, but not inescapably on the line y = x. Q – Q plots can also be used as a graphical means of estimating parameters in