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)

Answer: The effect of categorical variable on dependent variable can be explain as below.

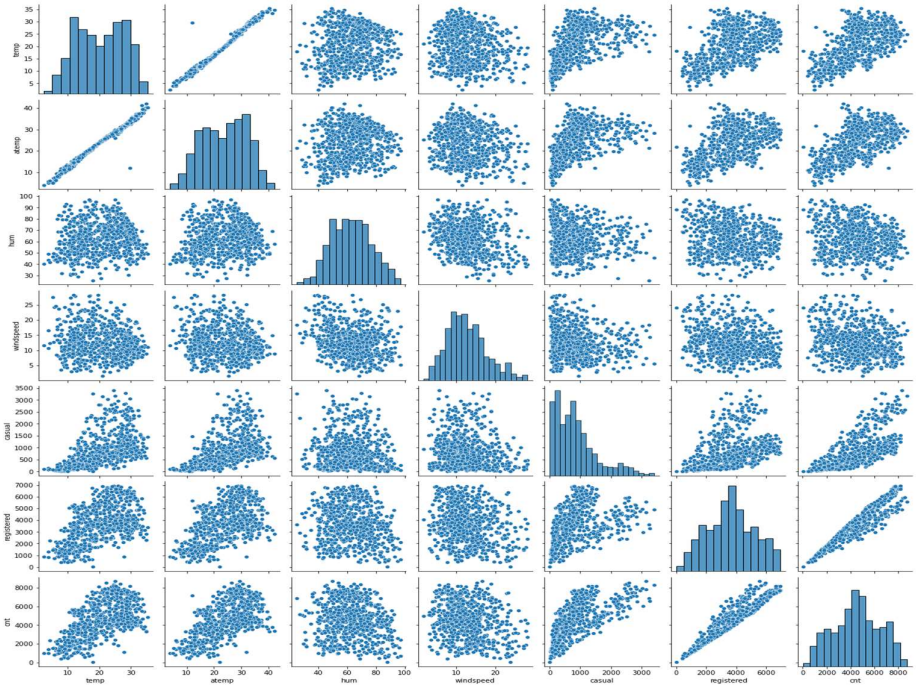
- The average bike rent is highest in season “fall” and lowest in “spring”. “Summer” and “winter” seems almost same trend and lower than “fall” but higher than “spring”
- There is a significant increment in average (overall bike) sales in 2019 as compare to 2018, so year seems a good predictor variable for bike rent forecasting.
- Weekday and Working day are shows almost similar trend on bike rent.
- There is a consistent increment in bike rent count from Jan to July then falls from July to Dec. probably it could be due to extreme weather condition during Dec.
- If the weather is clear then average bike rent is high compare to light snow rain and cloudy weather. Absolutely no rent when the rain is heavy.
- If Monday is working day there is absolutely no bike rent.

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

Answer: Dummy variables created N new feature variables and assigns either 0 or 1 where N is number categories in original categorical variable. The presence of one value means the absence of other values hence Drop_first=True means no assignment for first value, as it is by default explained by other new feature variables 0.

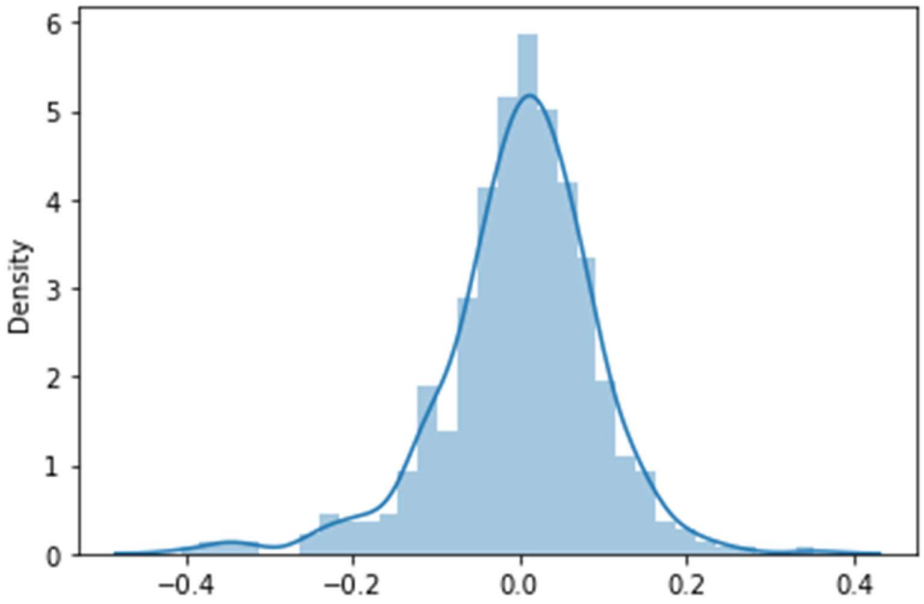
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer: As per pain plot the registered is having highest correlation with the target variable.



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

Answer: Draw the distribution plot for error term (predicted value – actual value) and it should follow normal distribution.



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

Answer: temp, humidity and year are top 3 features contributing significantly towards explaining the demand of the shared bikes.

General Subjective questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear Regression: A simple linear regression model attempts to explain the relationship between a dependent variable and an independent variable using a straight line. Below are some basic assumptions to fit a linear regression model for problem solving.

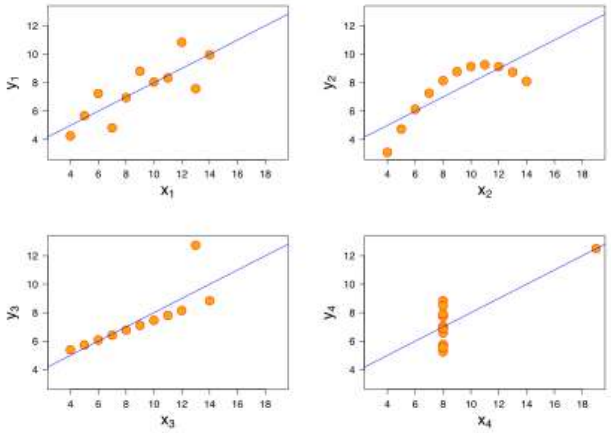
- There is a *linear relationship* between X and Y
- Error terms are *normally distributed* with mean zero(not X, Y)
- Error terms are *independent* of each other. I mean there should not be any visible pattern in error term.
- Error terms have *constant variance* (homoscedasticity. That means variance should not increase or decrease if error value changes.

R2 (correlation coefficient) is being used to visualize how much variance in dependent variable is being explained by independent variable.

MSE/RMSE/MAE are the error matrices used to determine the model's accuracy.

2. Explain the Anscombe's quartet in detail. 3 marks)

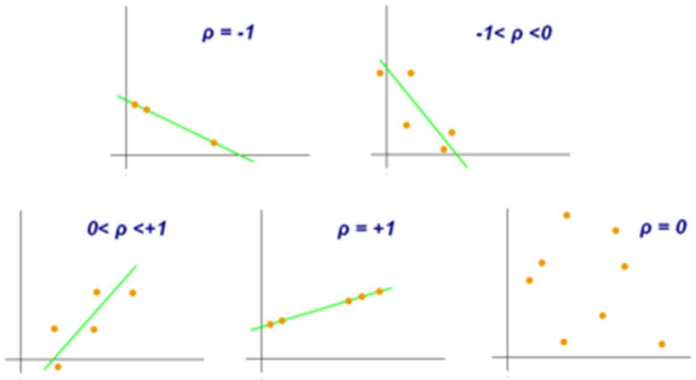
Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.



- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for).
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

3. What is Pearson's R? 3 marks)

Pearson coefficient or correlation coefficient R is a measure of linear relationship between 2 variables. It is the ratio between the covariance of two variables and the product of their standard deviations.



Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? 3 marks)

Scaling is to transform the value of any variable to some other range/scale. During machine learning model training our model may mistook a given high number feature variable is more significant than the low number feartue variable and hence give more weightage even though it is not that much significant.

Normalized Scaling: We also called it as normalization. Where we transform the value between 0 and 1 using max and min. hence we also called it as min-max scaling.

Xnew = (X – Xmin)/(Xmax – Xmin)

Standardized Scaling: we also called it standardization. Here we use mean and standard deviation (sigma) to rescale the value in the form of standard deviation (sigma)

Xnew = (X – mean)/SD(sigma)

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? 3 marks)

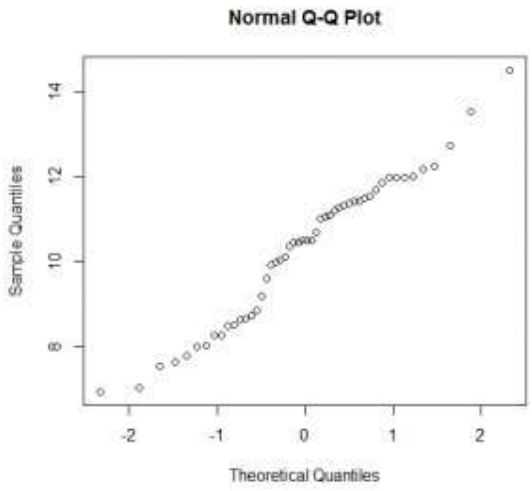
VIF = 1/(1-R2)

If R2 value becomes 1 somehow, then the value of VIF became infinite. Which is possible for highly correlated variable (having correlation coefficient 1). Normally this happens when same variable is being created with different names in dataset. E.g. temp (actual temperature)and atemp (feel like temperature)

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. 3 marks)

The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. For example, if we run a statistical analysis that assumes our residuals are normally distributed, we can use a Normal Q-Q plot to check that assumption. It's just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
8
10
12
14
-2
-1
0
1
2