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)**

   - There are no outliers in the categorical variables (one or two outliers maybe exceptional).
   - During Fall, the bike demand is very high, followed by Summer then Winter. Bike demand during Spring is the lowest. There is not much of difference in minimum & maximum number of bikes rented during all 4 seasons, with Spring season alone having a lower maximum rental than other seasons.
   - Every year, the bike demand follows a near normal distribution, with January and December being the lowest and, June and August being the highest, with a slight rise during September and October.
   - Bike demand during all the days of a week is nearly the same (50^{th} percentile is very similar, 75^{th} percentile is similar, min & max is almost similar).
   - Bike demand is the highest when there is a clear sky or a few clouds or is partly cloudy. Bike demand slightly reduces when there is some mist along with clouds or broken clouds or few clouds or just misty. Bike demand dips when there is drizzling with scattered clouds and thunderstorm or drizzling with only scattered clouds or just drizzling or falling of snowflakes. No bikes have been rented whenever there is a heavy downpour along with ice pallets, thunderstorm and mist or a snow fall with fog.
   - Bike demand has gone up considerably from the year 2018 to 2019 (75^{th} percentile during 2018 is nearly the same as 25^{th} percentile during 2019).
   - Sometimes, no bikes are rented during working days. On the other hand, there is not a single holiday when no bikes are rented. However, overall bike demand is high during working days. If the weekends are also considered as holidays, then the bike demand remains almost unaffected by whether or not a working day. Number of bikes rented during working days is nil, irrespective of this consideration.

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

   - If there are “n” categorical levels for a categorical variable then, there will be “n” dummy variables for it. However, “n - 1” dummy variables are enough to collectively represent the required information in the dataset.
   - This is because one dummy variable can be represented as a compliment of union of all other dummy variables.
   - Eliminating one of the dummy variables provides brevity to the information represented by each of the other dummy variables, including the one eliminated.
   - We use *drop_first=True* during dummy variable creation, as the first dummy variable has the Boolean value of $00…0_{n-1}$.

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

   - Both actual temperature and ‘feel-like’ temperature have the highest correlation with the target variable i.e., total count of bikes rented.
   - Both these numerical input variables are positively correlated with each other (very strong), as well as with the target variable (strong).
   - Humidity and windspeed are weekly correlated with the target variable.

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

- ✓ Validated if all the selected features are having low p-value and low VIF.
- ✓ Validated if the Model has zero mean Residue or no Error.
- ✓ Validated if the Model follows a normalized Error distribution with a constant Variance.
- ✓ Validated if the Model follows Homoscedasticity.
- ✓ Validated if the exhibits Linear behaviour i.e., each of the beta coefficients change by keeping other coefficients a constant.
- ✓ Validated if the Model does not exhibit Multicollinearity.
- ✓ Validated if the Model explains more than 80% of its Variance.
- ✓ Validated if the predictor variable explains more than 80% of the target variable's variance in the Model.
- ✓ Validated if the Model has a strong correlation between the predictors and the target variable and, compared the same between train & test sets.

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

Top 3 features influencing the demand of shared bikes are:

i. Change in temperature.
ii. Whether drizzling with scattered clouds and thunderstorm or drizzling with only scattered clouds or just drizzling or falling of snowflakes on that day.
iii. Change in humidity and windspeed.

General Subjective Questions

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

- Using linear regression algorithm, we try to predict the future trend by learning from the past and evaluating in the present.
- We may consider history (past) as sample data and the remaining (present and future) as population data.
- If we want to understand an element of the future, that is continuous in nature then, we perform regression.
- By using linear regression algorithm, we assume that the trend follows a linear pattern. Assumptions are as follows:
  a. change in input variable is directly proportional to the output variable.
  b. Error terms (residues) are normally distributed with its mean at zero.
  c. Error terms are independent of each other, unlike a time-series data.
  d. Error terms have constant variance.
  e. In case if there are more than one input variable then, we consider multicollinearity and variance inflation factor then, select the features accordingly, after scaling.
- Now, our model can be described using the equation of a straight line. We have the dependent variable (output) in the y-axis and, the independent variable (input) in the x-axis. There maybe one or more input variables but, only one output variable. We also have the straight-line intercept.
- We run a hypothesis test to check whether the input variable is significant in predicting the output variable or not, by knowing if the intercept is zero.
- We find the best fit line for our regression model using Ordinary Least Squares method. We obtain the best slopes and intercept from this line, by choosing the optimum p-values (also, variance inflation factor in case of multiple input variables).
- We validate all our assumptions of linear behaviour then, evaluate the predictability of our model.
- We ensure that our model can explain its variance significantly and, also if the predictor variable can explain target variable's variance significantly.
- Model is now ready for deployment.

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

- Anscombe's quartet consists of 4 completely different data points scatter pattern however, with nearly exact: mean of both input & output variables; variance of both input & output variables; correlation between input & output variables; linear regression (line of best fit); residual sum of squares.
- We can construct optimal statistics including linear regression parameters from simple descriptive statistics. We can also construct confidence intervals which quantify uncertainty of the parameter estimates. Anscombe formulated these quartets to stress the importance of data visualization.

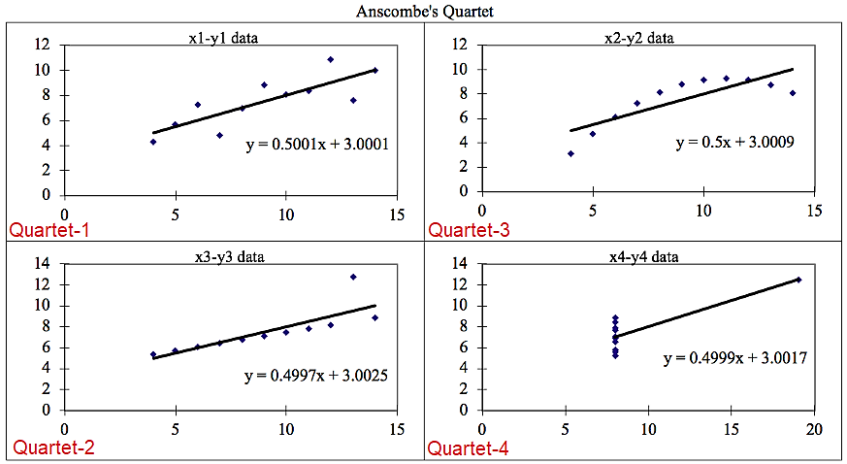

Image source: Towards Data Science Blog
Link: https://towardsdatascience.com/importance-of-data-visualization-anscombes-quartet-way-a325148b9fd2
*labelling done in red, on the extracted image, for ease of explanation

- We can obtain a best fit line, with a good RSS value from all the 4 quartets. In Quartet-1, there are no outliers and, hence no data imbalance. In Quartet-2, one outlier causes an imbalance in the average slope and intercept of the dataset. Quartet-4 has a linear relationship between input & output variables but, there are not many intermediate values to know if the fitted line is visually correct. Quartet-3 is not even linear.
- These quartets are the reason why we perform exploratory data analysis and, get insights from our data, before building our prediction model. We may need to remove the outliers, eliminate some variables, create new interactive features by combining two or more variables, before prediction.
- We need to know the importance of each data, regards to our problem statement and, Business requirements. After this, we perform EDA to check the variation of each variable in our dataset, both as individual (numerical & categorical) and, due to each other (numerical & categorical and, combined). Next, we need to clean our data set, retaining only relevant information then, perform a data sanity check, before model building.
- While model building, we need to ensure that we select the best set of features to have an optimum set of regression parameters (significance of best fit) for our model. Then, we also validate the assumptions of linearity on our model, before making any predictions.
- We balance between the Business requirements, influence of each data on the model, best fitting (linearity), significance of best fit, predictive power, optimum set of features, variance in data explained by both model for the entire data set and, by input variable for the output variable, in order to avoid misguidance of our prediction model, due to any of the Anscombe's quartets.

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

- Pearson's R or Pearson's correlation coefficient tells us how strongly or weekly any 2 variables are related/associated with each other.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Formula source: Byjus Learning Website
Link: https://byjus.com/pearson-correlation-formula/

where,

r = Pearson's R value
x = value of the data points at y=0
y = value of the data points at x=0
n = number of data points

- Pearson's R value is always between -1 and 1. When r = 1 then, both the variables vary exactly by the same units and, in same direction i.e., if one increases then, the other also increases. When r = -1 then, both the variables still vary exactly by the same units but, in opposite direction i.e., if one increases then, the other also decreases. When r = 0 then, the variables are not related to each other i.e., both randomly increases and/or decreases.
- Variables may have a strong (or) a weak correlation with a positive (or) a negative sign (or) no correlation at all.
  Pearson's R gives us both the magnitude and direction of their correlation.

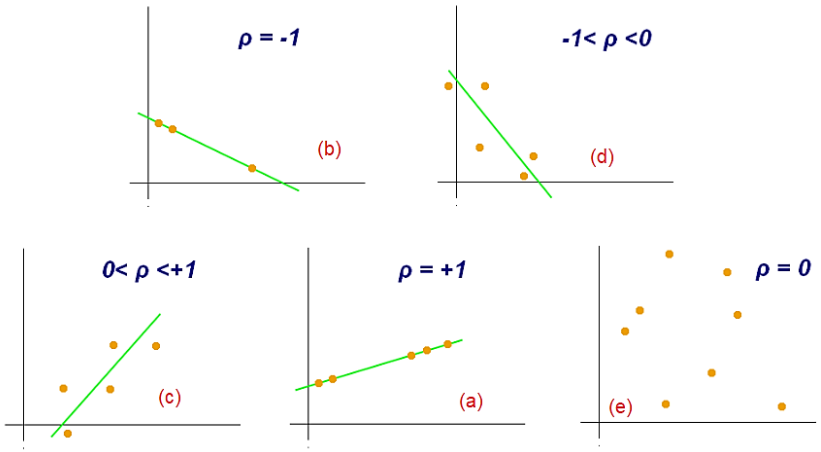

Image source: Byjus Learning Website
Link: https://byjus.com/correlation-coefficient-formula/
*labelling done in red, on the extracted image, for ease of reference

Based on this, Pearson's R can be classified (refer above image) as:

a. strong positive correlation
b. strong negative correlation
c. weak positive correlation
d. weak negative correlation
e. no correlation

- We can find the correlation between an input & output variables also, between any 2 input variables, using Pearson's R. We are able to effectively describe 'pair-wise' correlations quantitatively using Pearson's R and hence, also called as Pearson product-moment correlation coefficient.

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

- Definition and Formulae:

  Scaling is the process of normalizing the coefficients of all the predictor variables either in the standard normal distribution or min-max scale, thereby bringing all these coefficients within the same range.

  $$\text{Normalized Scaling} \Rightarrow X' = \frac{X - \bar{X}}{SD}$$

  $$\text{Standardized Scaling} \Rightarrow X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

  where,

  X = β coefficients or predictor coefficients
  X' = Scaled coefficient
  $\bar{X}$ = Mean coefficient
  SD = Standard Deviation
  X_{min} = Minimum coefficient's value
  X_{max} = Minimum coefficient's value

- Reason for Scaling:
  - When there are many predictor variables, generally in a multiple linear regression, sometimes even having interactive (combination of) features, each variable varies differently having different ranges. This makes it difficult to have a relative understanding of the variables.
  - When the predictors have many ranges then, its coefficients become less significant. If we get the features about the same ranges then, there will be a faster convergence of the gradients.
  - By doing this, we reduce the number of iterations, thus building a more meaningful model.
- Difference between normalized scaling and standardized scaling:

Sl. No.	Normalized Scaling	Standardized Scaling
1	Normalized based on values of the coefficients.	Normalized based on Z-score of the coefficients.
2	Outliers are taken care, so scaling is greatly affected by it.	Outliers are not taken care, so scaling is moderately affected by it.
3	Best for interpretation, rather than prediction.	Best for prediction, rather than interpretation.
4	Best if we are interested in the probability of a feature's trend.	Best if we are interested in the units of a feature's trend.
5	Coefficients of all the features follow the min-max scale.	Coefficients of all the features follow a standard normal distribution.
6	$0 < \beta_1, \beta_2, \ldots, \beta_p < 1$	$\bar{X} = 0$ & $SD = 1 \; \forall \; \beta_1, \beta_2, \ldots, \beta_p$

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

- Yes, I observed that the features 'Clear', 'Drizzle' and 'Misty' of the column "season" had infinite VIF values. VIF values of the features 'Drizzle' and 'Misty' came down to optimal levels when the feature 'Clear' was dropped.

$$VIF_i = \frac{1}{1 - R_i^2}$$

where,
VIF_i = Variance Inflation Factor for i^{th} input variable
R_i^2 = square of Residue of the i^{th} input variable

- Why sometimes $VIF_i \rightarrow \infty$?
  - Since, VIF α 1 / (1 - R^2), higher the value of R^2, lower is the value of (1 - R^2) and hence, higher the value of VIF.
  - If the VIF value is high, then higher is the association between the predictors since, higher R-squared value means, lower is the variance i.e., lower is the scatter with respect to the fitted line.
  - Each predictor variable is being represented as a linear combination of the rest of the independent variables. VIF value for i^{th} variable may change if we drop those predictors that are related to the i^{th} variable, if absolutely no association, then, VIF stays the same for that i^{th} variable.
  - Higher the VIF, higher the chances of one predictor variable being described by many others i.e., it is redundant in the presence of other variables.
  - Therefore, VIF becomes infinite when there is a perfect correlation i.e., that variable is being described by all (or) most of the other variables.

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

- Quartile divide the data points into 4 equal parts and, Percentile into 100 equal parts. Each of these equal parts is a Quantile. First quartile is the 25^{th} percentile, meaning that 25% of the total data points are below it, which is the first data segment. Second quartile is the 50^{th} percentile, meaning that 50% of the total data points are below it, which is the second data segment. Third quartile is the 75^{th} percentile, meaning that 75% of the total data points are below it, which is the third data segment. The remaining 25% of the data points above the third quantile, is the fourth quartile.
- A Quantile-Quantile plot or Q-Q plot tells us whether the 'n' number of data points present in our dataset are normally distributed or not.
- In order for us to have a Q-Q plot, we first divide our data into 'n' quantiles. Next, we draw a normally distributed curve defined by mean 'μ' and standard deviation 'σ', considering data points to be the random variable 'X'. From Inferential Statistics, we know that in a normally distributed curve:
  i. if $\mu - \sigma < X < \mu + \sigma$ then, P(X) = 68%
  ii. if $\mu - 2\sigma < X < \mu + 2\sigma$ then, P(X) = 95%
  iii. if $\mu - 3\sigma < X < \mu + 3\sigma$ then, P(X) = 99.7%
- Now, we divide the normally distributed curve into 'n' quantiles. Probability density is more towards the centre and less towards either of the ends. Since, we need to divide our data into 'n' spaces, each having an equal probability, we have narrow spaces towards the centre and, wider spaces towards either of the ends. By doing this, we ensure that the probability of observing a value is the same throughout the normal distribution.
- Finally, we draw a graph for the quantiles segmented from data points (according to range of data) versus the quantiles segmented from normally distributed curve (according to probability density), which is the Q-Q plot.

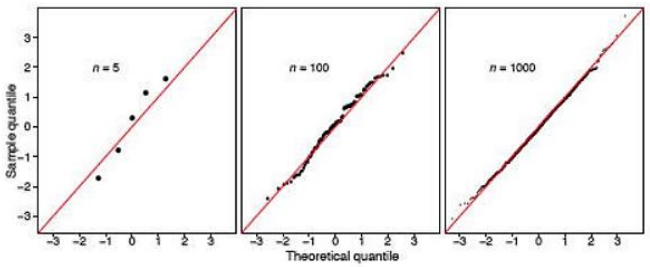

Figure 11. q-q plots of normal data.

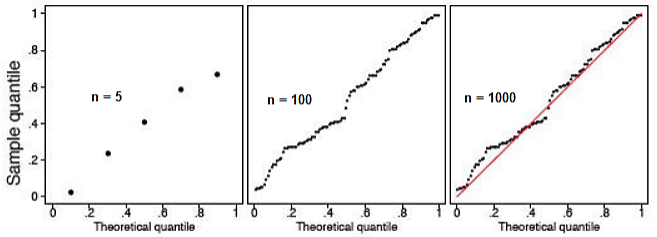

Figure 5. (Left) q-q plot of the 5 uniform points. (Right) q-q plot of a sample of 100 uniform points. and q-q plot of 1000 uniform points

Image source: Online Statistics Education: An Interactive Multimedia Course of Study, Rice University + University of Houston Clear Lake + Tufts University
Link: https://onlinestatbook.com/2/advanced_graphs/q-q_plots.html
*added the text “n = 5", “n = 100", “n = 1000", and “and q-q plot of 1000 uniform points” in Figure 5. for better understanding

- Once our Q-Q plot is done, we fit a straight line into the quantile-quantile points. If we observe that most of these points fall on the straight line then, our data points and the normally distributed curve have similar quantiles. In other words, we can say that our data points follow a normal distribution. On the other hand, if most of these points do not fall on the straight line then, we say that our data points are not normally distributed.
- In the latter case, we plot quantile-quantile points between the quantiles segmented from our data points (according to range of data) and a uniformly distributed graph (according to range of data). Note that a uniform distribution describes a step-wise straight line in the case of discrete random variables whereas, it describes a straight line, parallel to axis of variable ‘X’, in the case of continuous random variables.
- We fit a straight line into the quantile-quantile points in the Q-Q plot derived from uniform distribution, just as we did for the Q-Q plot derived from normal distribution. If we observe that most of these points fall on the straight line then, our data points and the uniformly distributed curve have similar quantiles. In other words, we can say that our data points follow a uniform distribution.
- We can also draw a Q-Q plot to find out whether a derived dataset is normally distributed or not, by comparing it with the actual dataset, the same way as we did by comparing a dataset with normal and/or uniform distribution(s). Here, we divide the original dataset i.e., large dataset into equal number of quantiles, same as that divided in the derived dataset i.e., much smaller dataset. This is because the smaller dataset cannot be divided further than a large dataset i.e., there will be more data segments below 75^{th} quantile (say) for the original dataset than that for the derived dataset.