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?

   Following **independent variables can have a bigger effect on the bike rentals**, as their distributions by category differ the most:

   - **weathersit**: Bike rentals will be done the most in clear with few clouds or partly cloudy days (69% of total for both 2018 and 2019 were done with those weather conditions).
   - **yr**: Bike rentals will increase every year. Compared to 2018, in 2019 bike rentals increased by 61.3%. Also, small amounts of bikes rental are rare in 2019.
   - **season**: Bike rentals will be done the most in fall and summer, and we can confirm that looking at the plots by month (mnth).
   - **mnth:** Bike rentals will be done the most in months corresponding to summer and fall, for days with no rain.

   Following **independent variables can influence** bike rentals but **not as significantly** as above ones:

   - **weekday**: Distribution of bike rentals is similar for all days of the week, but it will tend to be higher if the day is a Wednesday or a Saturday.
   - **holiday**: Amount of bike rentals done in days that did not correspond to a holiday is significantly bigger (90% for 2018 and 2019) than in holidays, so bike rentals will decrease in holidays.
   - **workingday**: Amount of bike rentals in not working days is significantly less (30% for 2018 and 2019) than not working days, so bike rentals will increase during working days.

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

   When we have dummy columns created for a variable with N different categories/levels, we can drop one column and still being able to explain all N levels. This means that if all N-1 categories' values are equal to zero, the record value corresponds to the *Nth* category that has been removed physically but not in meaning.

   For example, we have a variable called "genre" with possible values "female", "male", and "other".
   If we create dummy variables, we have possible records as in below table. If we delete "genre_other" column and we have zero value for both "genre_female" and "genre_male", we can still understand the *other* category has value 1.

genre_female	genre_male	~~genre_other~~
0	0	~~1~~
0	1	~~0~~
1	0	~~0~~

Table 1. Example of dummy variables for "genre" variable categories

The **parameter *drop_first* set to value *True*** indicates the function to drop the first dummy column created for a categorical variable, which as explained, will keep the explanation possible for all categories. This will allow us to **simplify the total number of columns used for our model**.

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

From below pair-plot among the numerical variables, we can see that **the one with highest correlation** with the target variable *cnt*, **is the independent variable *temp***.

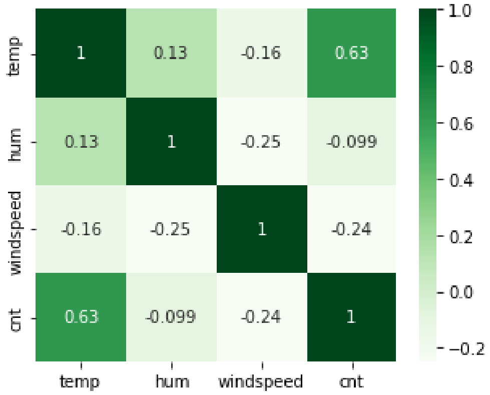


Figure 1. Correlation between numerical variables in Bike Sharing Demand case study

When p-value for this correlation coefficient was calculated it turned out to be zero. This means the correlation coefficient is significant and *temp* column has, indeed, a high correlation with target variable.

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

First, the **calculation of residuals** was done by computing the difference between original Y test values (target variable values) and predicted Y values. With these results we created a distribution plot to **check that we have the errors normally distributed**, with median and mean centered at value zero.

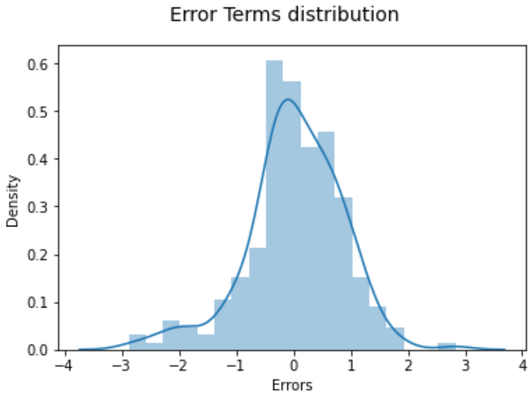


Figure 2. Distribution of error terms for the prediction set in Bike Sharing Demand case study

Then we plotted in a scatter plot the **distribution of predicted Y values against residuals**, to check that they do not follow any pattern, meaning they **are independent of each other**.

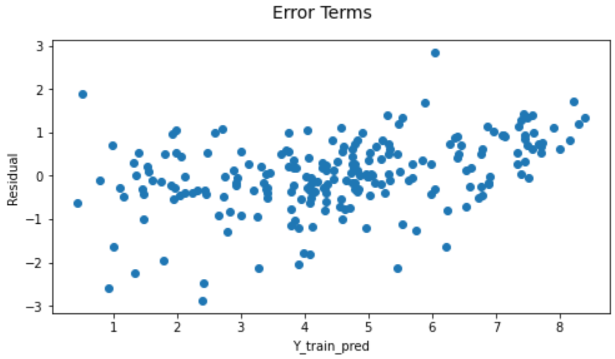


Figure 3. Distribution of predicted Y values against residuals in Bike Sharing Demand case study

Although there is still some unexplained **variance in the residuals**, it **does not increase** with Y values **or follow any pattern**. With this information we can rely in the significance of the model variables coefficients.

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

   These are the 3 variables that contribute the most to changes in bikes rental:

   - **temp** (positive relation)
   - **yr** (positive relation)
   - **weathersit_light_snowandrain** (negative relation)

   When the business team is planning the bikes supply, they need to take in account all variables found to be significant but give most consideration the 3 variables indicated above.
   The **best time to increase the supply** is during the **Summer**, when there are **high temperatures** but **not high windspeed or rain**.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Mathematically speaking, the linear regression method will find a **linear equation** in following form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

   Where:

   - y is the dependent variable.
   - β_0 is the line intercept.
   - $\beta_1 \ldots \beta_n$ are the coefficients of the independent variables, that determine the change by x_n unit in dependent variable, when all other variables are held constant.

As there are multiple options for each coefficient, the **criteria** to select the best line that describes the relationship is the one that **minimises the sum of the squares of the residuals**. This equation that describes how wrong the model is in finding a relation between the input and output is the **cost function.** For multiple linear regression, the cost function looks like the following:

$$J = \frac{1}{n}\sum_{i=0}^{n}(y_i - (\beta_i x_i + \beta_0))^2$$

Where:

- n is the total number of data points
- y_i is the actual value of an observation $\beta_i x_i + \beta_0$ is our prediction

The method to minimise the cost function is the **Ordinary Least Squares method,** which is done using the **Gradient descent optimization algorithm**. The Gradient descent algorithm has an input parameter called **learning rate**, a scale factor for selecting the values of the coefficients.

The Gradient descent algorithm steps are as follows:

a. Select first random values for each variable coefficient.
b. The sum of the squared errors is determined for each pair of input and output values.
c. The coefficients are updated in the direction towards minimising the error, using the learning rate.
d. The process is repetitive: until a minimum sum squared error has been reached or no further improvement is possible the algorithm stops.

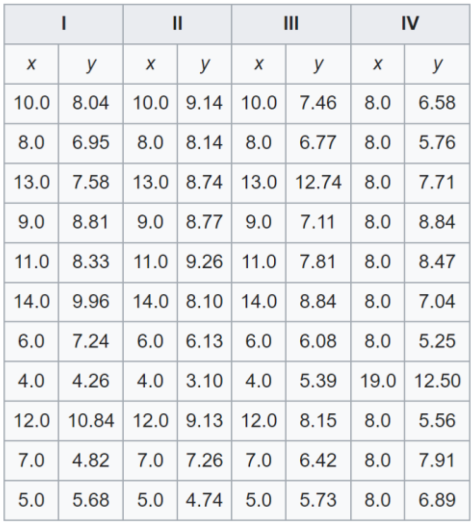
2. Explain the Anscombe's quartet in detail.

Anscombe's quartet consist of **four data sets that have the same statistical properties but** are obviously **different** when inspecting their respective **graphs**.

For each data set we have with high accuracy:

- Mean of x: 9
- Sample variance of x: 11
- Mean of y: 7.50
- Sample variance of y: 4.125
- Correlation between x and v: 0.816
- Linear regression line: $y = 3 + 0.5x$
- R-squared: 0.67

Table at the right shows the four data sets, and image in following page shows the graph for each data set.



I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Table 2. Anscombe's quartet datasets
https://en.wikipedia.org/wiki/Anscombe%27s_quartet

The Anscombe's quartet is used to emphasize the **importance of visualizing the data before analyzing it**, as the distribution may be describing very different scenarios even if the statistical properties are equal.

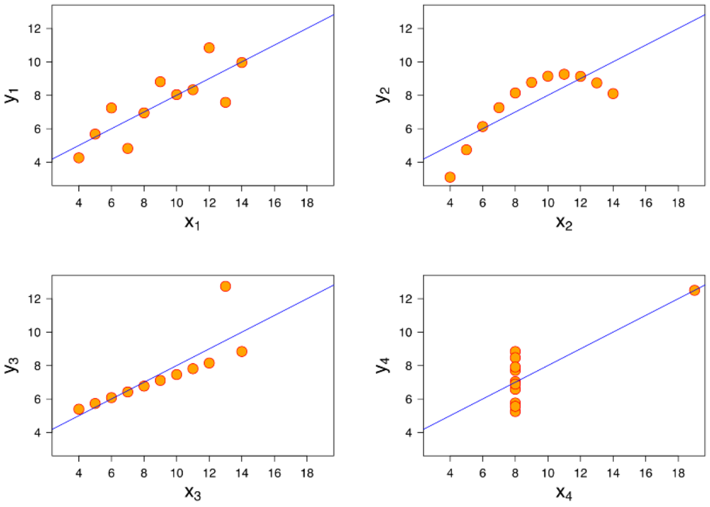


Figure 4. Anscombe's quartet distributions [https://en.wikipedia.org/wiki/Anscombe%27s_quartet#/media/File:Anscombe's_quartet_3.svg]

3. What is Pearson's R?

To describe the relationship between two variables we need to be able to describe certain aspects:

- If changes in the value of one of the variables come with changes in the value of the other variable.
- If the variables appear to be related, are they closely or only slightly related?
- If the variables appear to be related, is the relationship positive or negative?

Pearson's R allows us to explain the aspects above, as it is a **numerical measure of the correlation between two variables**.

Pearson's R is represented on a scale of 0 to 1, in both the positive and negative directions. A value of "0" indicates that there is no linear relationship between the variables. A value of "1" or "–1" indicates, respectively, a perfect positive or perfect negative correlation between two variables.

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

Scaling is the transformation of the distribution of a variable. This transformation is performed with the **purpose** of being able to **make comparisons** with elements of **different variables** and **different units** of measurement, so the entire data set will be unitless.

When scaling data for Machine Learning, it also helps improving the different methods performance, as small and fixed ranges of values are used.

There are two main methods of scaling, *normalization,* and *standardization*.

Normalization transforms the values of a variable into a range between 0 and 1, it reduces the effects of outliers in data.

Standardization transforms the values of a variable such that the resulting distribution has a mean of 0 and a standard deviation of 1.

In the following image, you can see the difference between normalization and standardization for a variable of temperature called "temp". The values range differ as well as the median and mean.

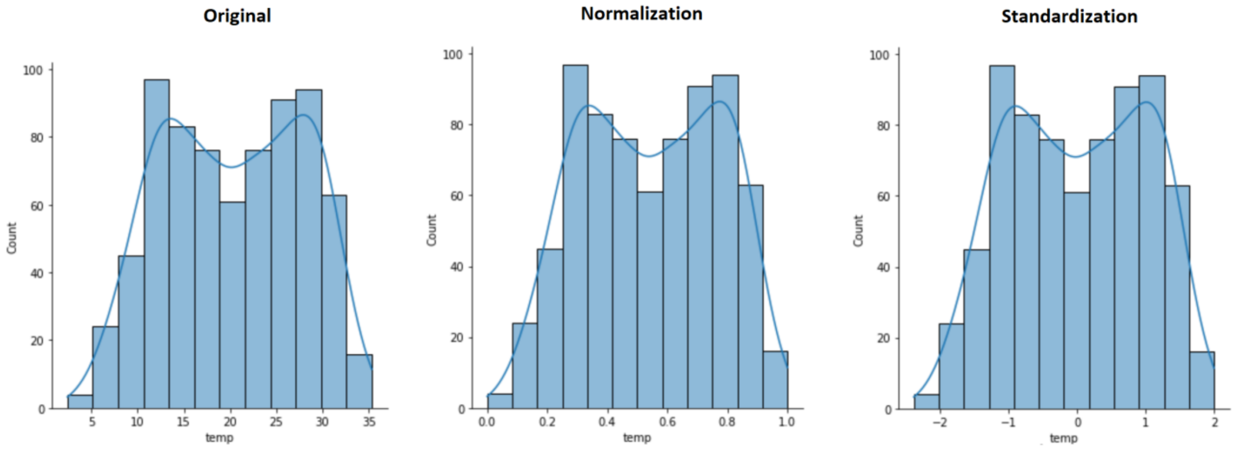


Figure 5. Normalization vs. Standardization of a variable of temperature

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

The **Variance Inflation Factor (VIF)** is mathematically defined for a i-th variable in a linear equation as:

$$VIF_i = \frac{1}{1 - R^2{}_i}$$

To obtain a **VIF** equal to **infinity**, $R^2{}_i$ value must be equal to 1. A value of $\boldsymbol{R^2{}_i}$ **equal to 1** means that the **dependent variable** can be **perfectly explained** by a **linear function of the i-th variable**. Following image represents this scenario.

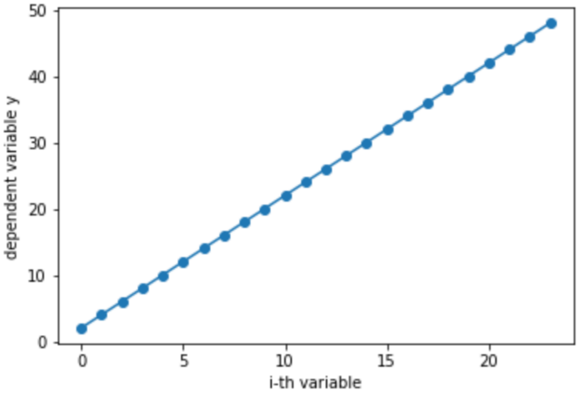


Figure 6. Perfect linear relationship between i-th variable and dependent variable y

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

A **Q-Q plot**, or **Quantile-Quantile plot**, allows you to see how close the distribution of a data set is to some ideal distribution, or **compare** the **distribution** of two data sets. It is called like that as the quantiles of two variables are plotted against each other.

An example is shown in image below, where we compare a sample of data on the vertical axis to a statistical population on the horizontal axis. The points are lying nearly on the straight line so we can say the distributions are similar.

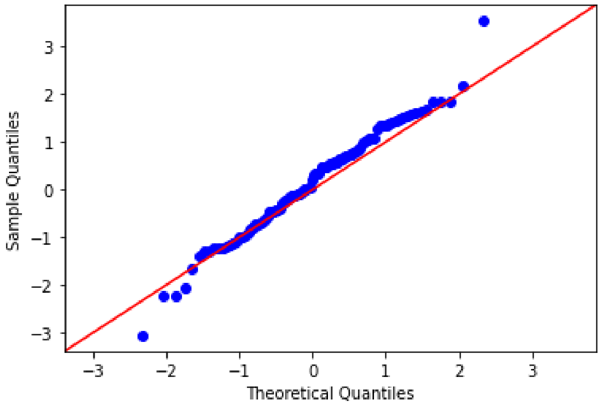


Figure 7. A normal Q-Q plot of randomly generated data (generated with StatsModel library in Python)

This is useful in **linear regression** when we want to compare **train and test sets** that came from different sources. Using Q-Q plot we can confirm that both the data sets are from populations with **same distributions** and we can continue on modeling with both of them.