Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

From our dataset, the key categorical variables are:

- **Season (season)**: Spring, Summer, Fall, Winter
- **Weather Situation (weathersit)**: Clear, Mist/Cloudy, Light Snow/Rain, Heavy Rain/Thunderstorm
- **Month (mnth)**: January to December
- **Weekday (weekday)**: Monday to Sunday
- **Holiday (holiday)**: Whether the day is a holiday or not
- **Working Day (workingday)**: Whether the day is a working day

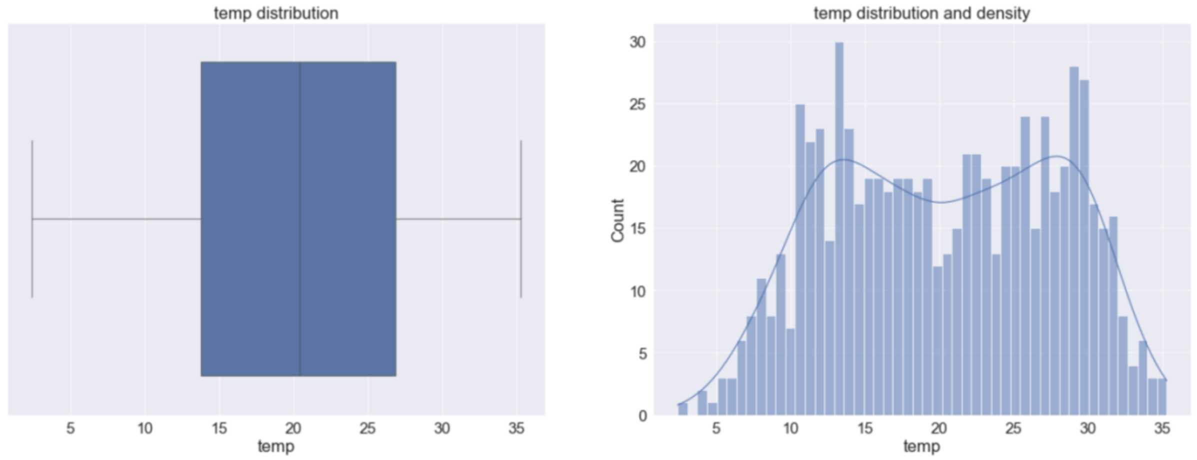
Season and Month Influence Bike Demand

Inference: Seasonality affects bike demand, with peak rentals in Fall and Summer and lower demand in Winter.

Weathersit (Weather Conditions) Affects Demand Significantly

- **Clear weather (weathersit=1) leads to high demand.**
- **Rainy/snowy days (weathersit=3 or 4) see a sharp drop in demand.**

Inference: Bad weather negatively impacts bike rentals, as people avoid cycling in heavy rain, thunderstorms, or snow.



Weekday and Workingday Have an Interesting Pattern

- Weekdays (**workingday=1**) generally have higher demand from registered users.
- Weekends (**workingday=0**) may see higher casual user demand but lower total rentals compared to weekdays.
- Certain weekdays (like Friday) might show a slight increase due to early weekend usage.

Inference:

1. **Commuters use bikes more on working days**, leading to higher rentals from registered users.
2. **Casual riders increase on weekends, but total rentals may not always be higher.**

Holiday Lowers Overall Bike Demand

- **On holidays (holiday=1), total rentals tend to drop.**
- **Registered users (commuters) contribute significantly to overall demand, so a lack of work-related trips on holidays reduces demand.**

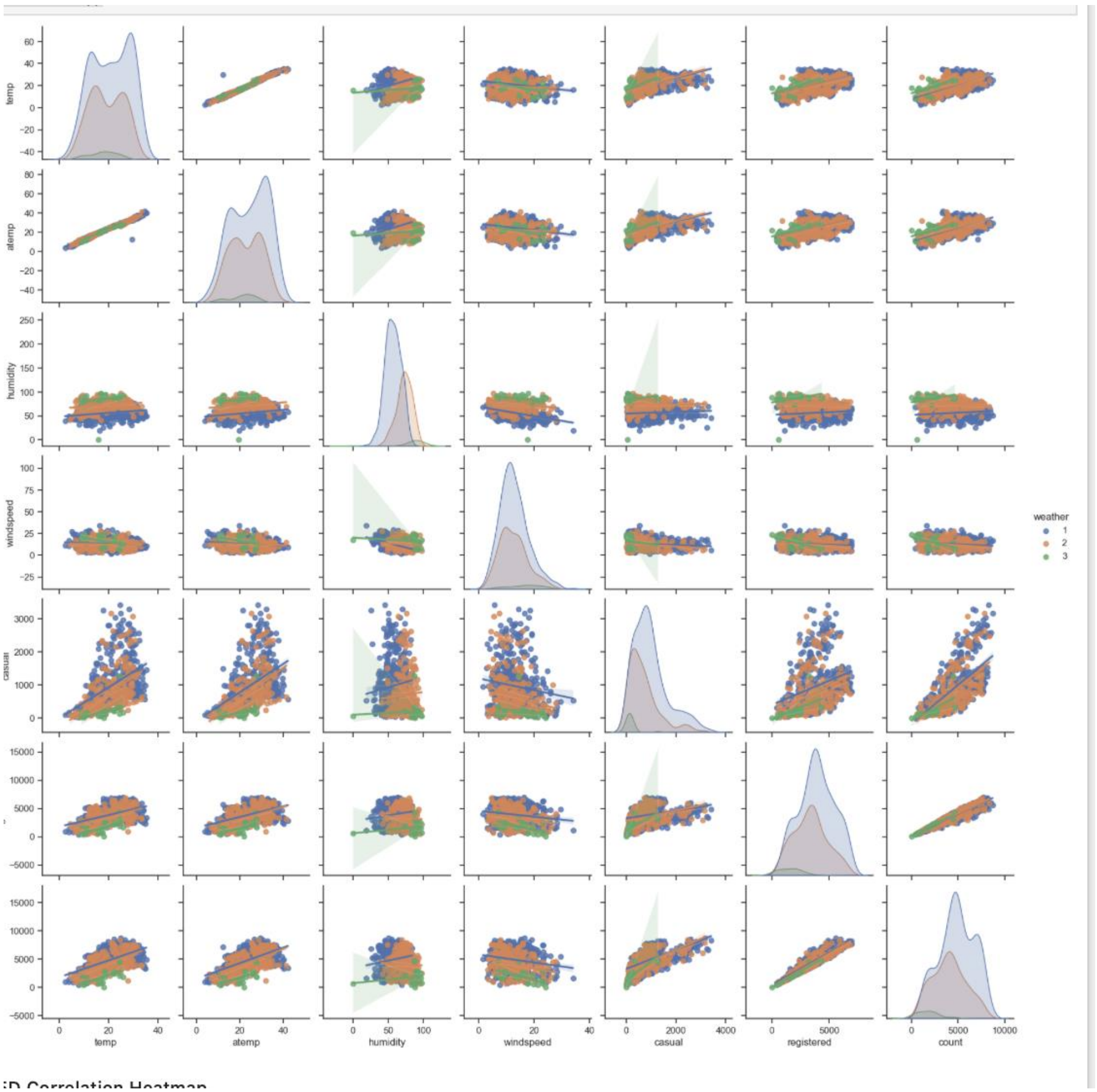
Inference: Bike rentals drop on holidays, as fewer people commute.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 2 goes below this line> (Do not edit)

Dummy variables are created when there is categorical variables involved with various options. It is essential to to use drop_first = true for various reasons,

1. **Reducing Features**:
   - Using drop_first=True reduces the number of features by one. This helps with simplifying the model while retaining the full information.
2. **Remove Multicollinearity**:
   - Multicollinearity is caused in linear regression when independent variables are highly correlated. When dummy variables are created and drop_first is not used, it can lead to having one of the variable to be highly correlated with the rest.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)

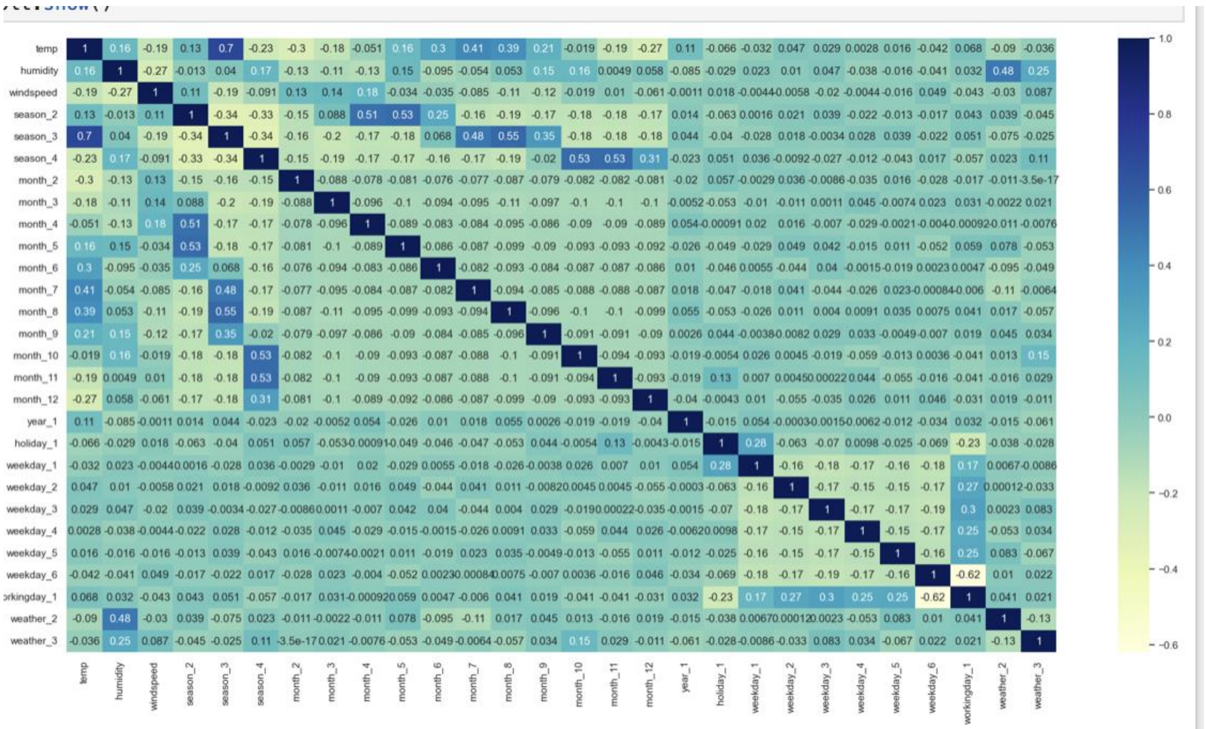


Temp and atemp have a good positive correlation with count variable. We are not considering registered and casual since they are also part of the dependent variable.

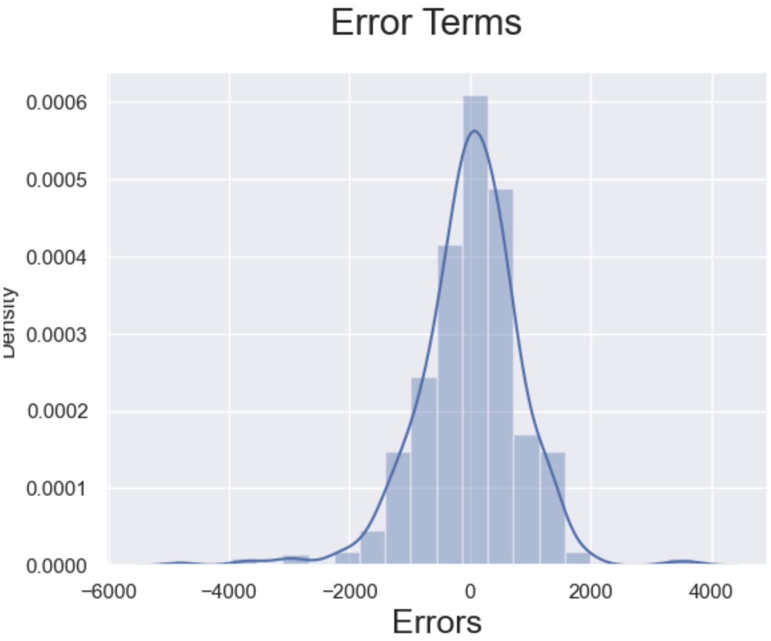
Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

Assumptions on Linear regression:
It is assumed that there is a linear relationship between the dependent and independent variables. It is known as the ‘linearity assumption’.

The scatter plots help identify the linear relationship between the dependent and independent variables. I have used the pair plots, heatmaps and correlation matrices to look at the relationships between given variables.



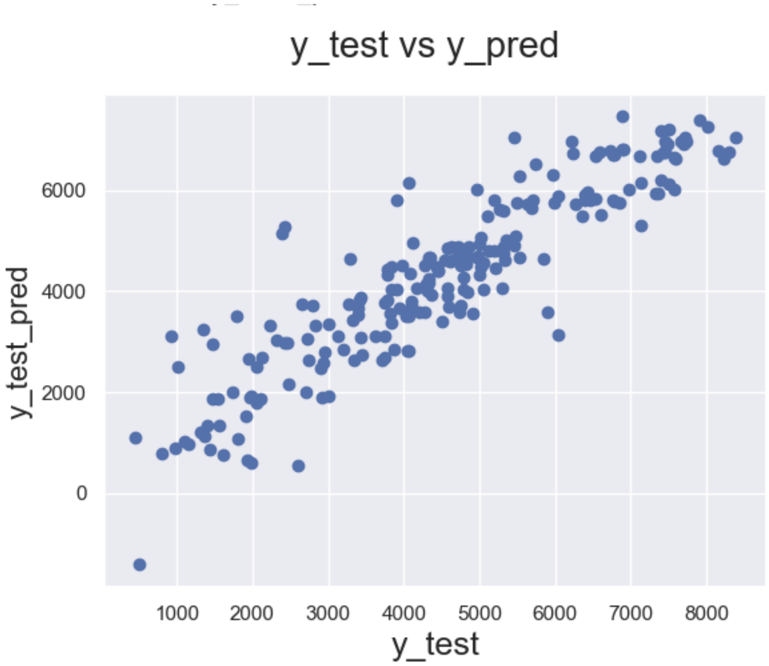
Assumptions about the residuals:
It is assumed that the error terms, are normally distributed. By reviewing normal distribution plots on the residuals(y train , y predict) we can see if they are normally distributed.



Zero mean assumption:

It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero. This can be verified from the above plot.

It is assumed that the residual terms have the same (but unknown) variance, σ2. This assumption is also known as the assumption of homogeneity or homoscedasticity. The scatter plot relative to Residual vs Y_test_predict can help verify this.



It is assumed that the residual terms are independent of each other, i.e., their pair-wise covariance is zero. The scatter plot relative to Residual vs Y_test_predict helps verify the same.

Assumptions about the estimators:

• The independent variables are measured without error.

• The independent variables are linearly independent of each other, i.e., there is no multicollinearity in the data. – have used the VIF factor to verify the multicollinearity in the data and dropping the variables accordingly to reduce the same and increase model significance.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 5 goes below this line> (Do not edit)

Major drivers impacting demand are:

* Scaling demand between years 1 & 2

* Top 3 positively weighted variables are:

1. Year
2. temp
3. winter
* Top 3 negatively weighted variables are:
1. Weather: Light Snow, Light Rain
2. Windspeed
3. Weather: Mist + Cloudy, Mist

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Linear Regression Algorithm: A Detailed Explanation

In simple terms, linear regression is a method of finding the best straight-line fitting to the given data, i.e., finding the best linear relationship between the independent and dependent variables.

In technical terms, linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables. It is mostly done by the Residual Sum of Squares Method.

Assumptions in a linear regression model:

Assumption about the form of the model: It is assumed that there is a linear relationship between the dependent and independent variables. It is known as the 'linearity assumption'.
Assumptions about the residuals:

- Normality assumption: It is assumed that the error terms, ε(i), are normally distributed.
- Zero mean assumption: It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.
- Constant variance assumption: It is assumed that the residual terms have the same (but unknown) variance, σ2. This assumption is also known as the assumption of homogeneity or homoscedasticity.
- Independent error assumption: It is assumed that the residual terms are independent of each other, i.e., their pair-wise covariance is zero.

Assumptions about the estimators:

- The independent variables are measured without error.
- The independent variables are linearly independent of each other, i.e., there is no multicollinearity in the data.

Mathematically, we can write a linear regression equation as:

$Y = a + bx$

a,b are y-intercept and slope of the line.

Use Cases of Linear Regression:

1. Prediction of trends and Sales targets – To predict how industry is performing or how many sales targets industry may achieve in the future.

2. Price Prediction – Using regression to predict the change in price of stock or product.

3. Risk Management- Using regression to the analysis of Risk Management in the financial and insurance sector.

Hypothesis testing in linear regression

Hypothesis testing can be carried out in linear regression for the following purposes:

1. To check whether a predictor is significant for the prediction of the target variable. Two common methods for this are as follows:

    a. Using p-values: If the p-value of a variable is greater than a certain limit (usually 0.05), the variable is insignificant in the prediction of the target variable.

    b. By checking the values of the regression coefficient: If the value of the regression coefficient corresponding to a predictor is zero, that variable is insignificant in the prediction of the target variable and has no linear relationship with it.

2. To check whether the calculated regression coefficients are good estimators of the actual coefficients. The null and alternative hypotheses used in the case of linear regression, respectively, are:

$\beta 1=0$
$\beta 1\neq 0$

Thus, if we reject the null hypothesis, we can say that the coefficient β1 is not equal to zero and, hence, is significant for the model. On the other hand, if we fail to reject the null hypothesis, we can conclude that the coefficient is insignificant and should be dropped from the model.

Interpretation of a linear regression model:

A linear regression model is quite easy to interpret. The model is of the following form:

$y=\beta 0+\beta 1X1+\beta 2X2+...+\beta nXn$

The significance of this model lies in the fact that one can easily interpret and understand the marginal changes and their consequences. For example, if the value of x_0 increases by 1 unit, keeping other variables constant, the total increase in the value of y will be β_i. Mathematically, the intercept term (β_0) is the response when all the predictor terms are set to
zero or not considered.

Shortcomings of linear regression:
You should never just run a regression without having a good look at your data because simple linear regression has quite a few shortcomings:

- It is sensitive to outliers.
- It models linear relationships only.
- A few assumptions are required to make the inference.

Parameters used to check the significance of the model and the goodness of fit?

- **F-statistics** – To check whether the overall model fit is significant or not, the primary parameter to be looked at is the F-statistic.
- **T-test**, along with the p-values for betas, tests whether each coefficient is significant or not individually.
- **R-squared** (for simple linear regression models) or adjusted R-squared (for multiple linear regression models). If your overall model fit is deemed to be significant by the F-test, you can go ahead and look at the value of R-squared. This value lies between 0 and 1, with 1 meaning a perfect fit. A higher value of R-squared is indicative of the model being good with much of the variance in the data being explained by the straight line fitted. For example, an R-squared value of 0.75 means that 75% of the variance in the data is being explained by the model. But it is

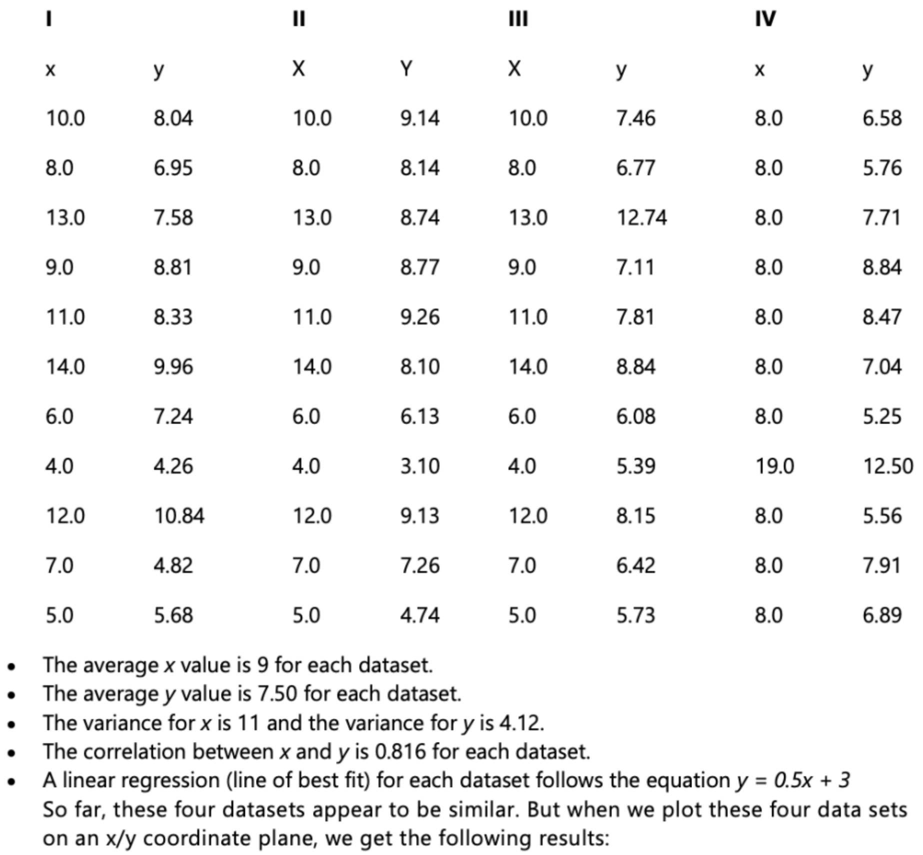
important to remember that R-squared only tells the extent of the fit and should not be used to determine whether the model fit is significant or not.

Ref: https://www.springboard.com/blog/data-science/what-is-linear-regression/

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Perhaps the most elegant demonstration of the dangers of summary statistics is Anscombe's Quartet. It is a group of four datasets that appear to be similar when using typical summary statistics yet tell four different stories when graphed. Each dataset consists of eleven (x,y) pairs as follows:

A Real-World Example
Let us look at a real dataset that shows exactly how summary statistics can be dangerous. A great example is the distribution of starting salaries for new law graduates. The National Association of Law Placement (NALP) reports that in 2012, lawyers made $80,798 on average in starting salary. However, a look at the salary distribution shows what law salaries really look like:



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

- The average *x* value is 9 for each dataset.
- The average *y* value is 7.50 for each dataset.
- The variance for *x* is 11 and the variance for *y* is 4.12.
- The correlation between *x* and *y* is 0.816 for each dataset.
- A linear regression (line of best fit) for each dataset follows the equation *y = 0.5x + 3*
  So far, these four datasets appear to be similar. But when we plot these four data sets on an x/y coordinate plane, we get the following results:

It turns out that law graduates usually fall into one of two groups. Most new lawyers make somewhere between $35,000 and $75,000 per year, and a sizable minority earns

$160,000 per year. What we have here is a bimodal distribution: there are two peaks that arise from two distinct distributions happening within the same dataset. The $80,798 figure reported as the average falls into the trough between the two peaks, and few lawyers have salaries near that number. A much more accurate statement would be that most law graduates make around $50,000 on average, and those who go to one of the top law schools make $160,000 on average.

There is also something else happening here that we would not have observed if we had not plotted the data. There is a giant spike at exactly $160,000 in starting salary, rather than a peak with some variance. Why is $160,000 such a popular number for law salaries? As it turns out, this data is not based on actual legal salaries, but based on what law schools report to the NALP as their students' median starting salaries. There is a lot of scepticism about the $160,000 figure, and third-party data shows that the distribution might not be so skewed.

Visualizing the data helped in two ways. It gave us a better picture of what realistic starting law salaries look like and allowed us to ask a follow-up question that exposed a potential flaw in our data.

When should you use summary statistics?

This is not to say that summary statistics are useless. They are just misleading on their own. It is important to use these as just one tool in a larger data analysis process. Visualizing our data allows us to revisit our summary statistics and recontextualize them as needed. For example, Dataset II from Anscombe's Quartet demonstrates a strong relationship between x and y, it just doesn't appear to be linear. So a linear regression was the wrong tool to use there, and we can try other regressions. Eventually, we will be able to revise this into a model that does a great job of describing our data and has a high degree of predictive power for future observations.

Ref: https://heap.io/blog/data-stories/anscombes-quartet-and-why-summary-statistics-dont-tell-the-whole-story

---

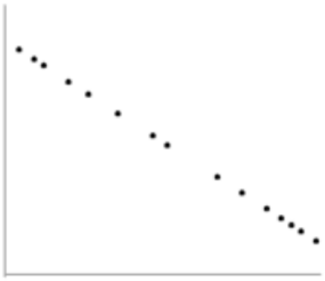
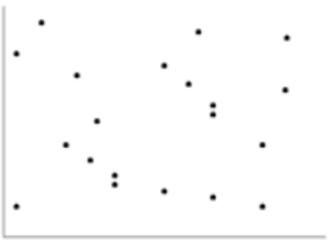
Question 8. What is Pearson's R? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

Pearson's Correlation Coefficient is a technique for investigating the relationship between two quantitative, continuous variables, for example, age and blood pressure. Pearson's correlation coefficient (r) is a measure of the strength of the association between the two variables. The first step in studying the relationship between two continuous variables is to draw a scatter plot of the variables to check for linearity. The correlation coefficient should not be calculated if the relationship is not linear. For correlation only purposes, it does not really matter on which axis the variables are plotted. However, conventionally, the independent (or explanatory) variable is plotted on the x-axis (horizontally) and the dependent (or response) variable is plotted on the y-axis (vertically).

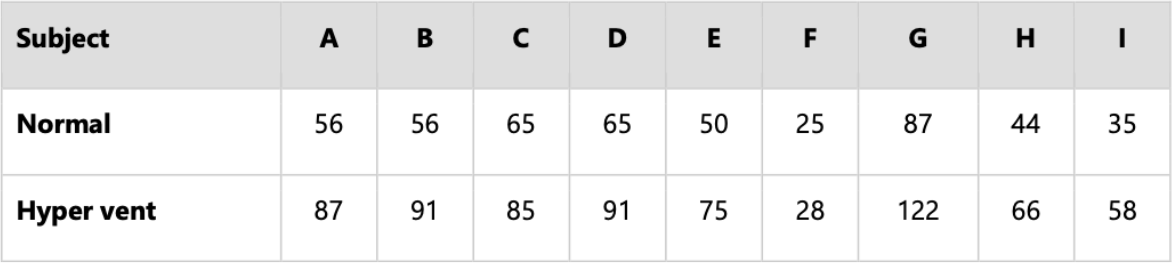
The nearer the scatter of points is to a straight line, the higher the strength of association between the variables. Also, it does not matter what measurement units are used. Pearson's correlation coefficient (r) for continuous (interval level) data ranges from -1 to +1:

r = -1		data lie on a perfect straight line with a negative slope
r = 0		no linear relationship between the variables

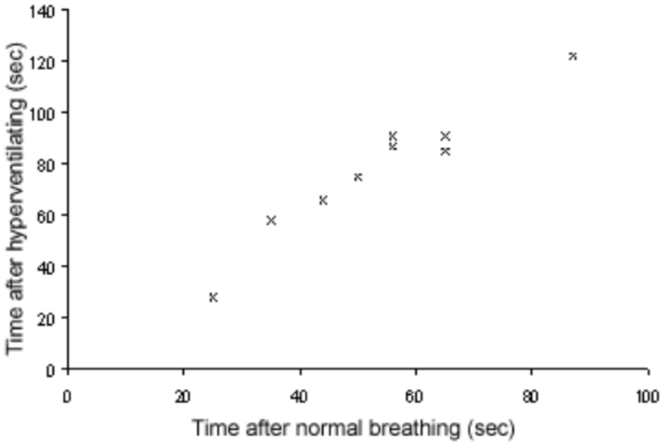
---

Positive correlation indicates that both variables increase or decrease together, whereas negative correlation indicates that as one variable increases, so the other decreases, and vice versa.
Example : Scatterplots Identify the approximate value of Pearson's correlation coefficient. There are 8 charts, and on choosing the correct answer, you will automatically move onto the next chart.

Tip: that the square of the correlation coefficient indicates the proportion of variation of one variable 'explained' by the other (see Campbell & Machin, 1999 for more details).
Statistical significance of r Significance. The t-test is used to establish if the correlation coefficient is significantly different from zero, and hence that there is evidence of an association between the two variables. There is then the underlying assumption that the data is from a normal distribution sampled randomly. If this is not true, the conclusions may well be invalidated. If this is the case, then it is better to use Spearman's coefficient of rank correlation (for non-parametric variables). See Campbell & Machin (1999) appendix A12 for calculations and more discussion of this. It is interesting to note that with larger samples, a low strength of correlation, for example $r = 0.3$, can be highly statistically significant (i.e., $p < 0.01$). However, is this an indication of a meaningful strength of association?
NB Just because two variables are related, it does not necessarily mean that one directly causes the other!
Worked example :Nine students held their breath, once after breathing normally and relaxing for one minute, and once after hyperventilating for one minute. The table indicates how long (in sec) they were able to hold their breath. Is there an association between the two variables?

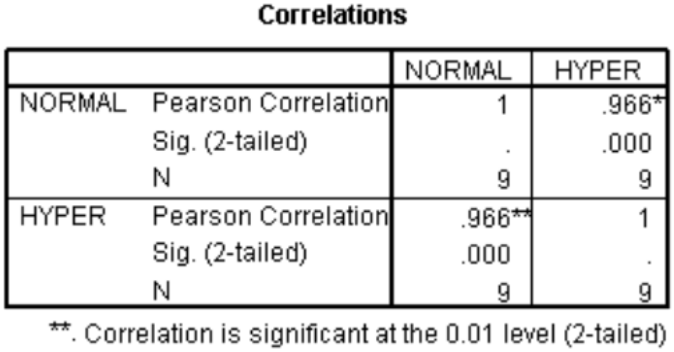


Subject	A	B	C	D	E	F	G	H	I
Normal	56	56	65	65	50	25	87	44	35
Hyper vent	87	91	85	91	75	28	122	66	58



The chart shows the scatter plot (drawn in MS Excel) of the data, indicating the reasonableness of assuming a linear association between the variables. Hyperventilating times are the dependent variable, so are plotted on the vertical axis.

Output from SPSS and Minitab are shown below:
SPSS
Select Analysis>Correlation>Bi-variate



Correlations

		NORMAL	HYPER
NORMAL	Pearson Correlation	1	.966*
	Sig. (2-tailed)	.	.000
	N	9	9
HYPER	Pearson Correlation	.966**	1
	Sig. (2-tailed)	.000	.
	N	9	9

**. Correlation is significant at the 0.01 level (2-tailed)

Pearson correlation of Normal and Hyper vent = 0.966
P-Value = 0.000
In conclusion, the printouts indicate that the strength of association between the variables is very high (r = 0.966), and that the correlation coefficient is very highly significantly differentfrom zero (P < 0.001). Also, we can say that 93% (0.9662) of the variation in hyperventilating
times is explained by normal breathing times.

Ref: http://learntech.uwe.ac.uk/da/Default.aspx?pageid=1442

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. ... If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.
Standardized scaling:
This means that you are transforming your data so that it fits within a specific scale, like 0–100 or 0–1. You want to scale data when you're using methods based on measures of how far apart data points, like support vector machines, or SVM or k-nearest neighbours, or KNN. With these algorithms, a change of "1" in any numeric feature is given the same importance. By scaling your variables, you can help compare different variables on equal footing.
Normalization Scaling just changes the range of your data. Normalization is a more radical transformation. The point of normalization is to change your observations so that they can be described as a normal distribution.
Normal distribution: Also known as the "bell curve", this is a specific statistical distribution where a roughly equal observations fall above and below the mean, the mean and the median are the same, and there are more observations closer to the mean. The normal distribution is also known as the Gaussian distribution.
In general, you will only want to normalize your data if you are going to be using a machine learning or statistics technique that assumes your data is normally distributed. Some examples of these include t-tests, ANOVAs, linear regression, linear discriminant analysis (LDA) and Gaussian naive Bayes. (Pro tip: any method with "Gaussian" in the name probably assumes normality.)

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

VIF: VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. To determine VIF, we fit a regression model
between the independent variables. For example, we would fit the following models to estimate the coefficient of determination R1 and use this value to estimate the VIF:
If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect
correlation, then VIF = infinity.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
(Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Quantile-Quantile (Q-Q) plot is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential, or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.
This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.
Few advantages:
a) It can be used with sample sizes also.
b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and
the presence of outliers can all be detected from this plot.
It is used to check following scenarios:
If two data sets —
i. come from populations with a common distribution.
ii. have common location and scale.
iii. have similar distributional shapes.
iv. have similar tail behaviour.
Interpretation:
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.