1. What are the key tasks involved in getting ready to work with machine learning modeling?

* **Clean the data**
* **Remove outliers/deal with them**
* **Remove or replace null values**
* **Split the data into train test split**
* **Decide suitable model**
* What are the different forms of data used in machine learning? Give a specific example for each of them.
  + **Training data**
  + **Validation data**
  + **Test data**

3. Distinguish:

1. Numeric vs. categorical attributes

|  |  |
| --- | --- |
|  | |
|  | **Categorical data refers to a data type that can be stored and identified based on the names or labels given to them.** | | **Numerical data refers to the data that is in the form of numbers, and not in any language or descriptive form.** |
|  | **Also known as qualitative data as it qualifies data before classifying it.** | | **Also known as quantitative data as it represents quantitative values to perform arithmetic operations on them.** |
|  | **What is your gender?**   * **Male** * **Female** * **Other** | | **What is your test score out of 20?**   * **Below 5** * **5-10** * **10-15** * **15-20** * **20** |
|  | **Nominal data and Ordinal data.** | | **Discrete data and Continuous data.** |
|  | * **No order scale** * **Natural language description** * **Can take numerical values but with qualitative properties** * **Can be visualized using bar charts and pie charts** | | * **Has an ordered scale** * **Not use of natural language description** * **Takes numeric values with numeric qualities** * **Can be visualized using bar charts and pie charts** |
|  | **Can include long surveys and has a chance of pushing respondents away.** | | **Survey interaction is easy and short, hence fewer survey abandonment issues.** |

2. Feature selection vs. dimensionality reduction

4. Make quick notes on any two of the following:

**the histogram**

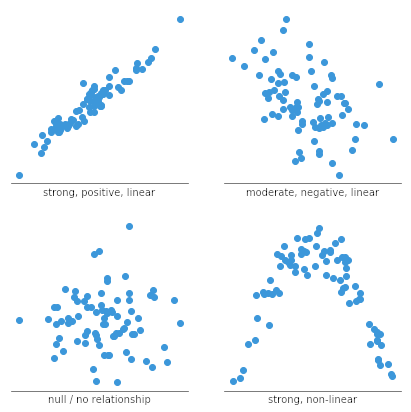
* **A frequency distribution shows how often each different value in a set of data occurs. A histogram is the most commonly used graph to show frequency distributions. It looks very much like a bar chart.**

**Use a histogram when:**

* + **The data are numerical**
  + **You want to see the shape of the data’s distribution, especially when determining whether the output of a process is distributed approximately normally**
  + **Analyzing whether a process can meet the customer’s requirements**
  + **Analyzing what the output from a supplier’s process looks like**
  + **Seeing whether a process change has occurred from one time period to another**

1. Use a scatter plot

* **A scatter plot (aka scatter chart, scatter graph) uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. Scatter plots are used to observe relationships between variables.**
* **Scatter plots’ primary uses are to observe and show relationships between two numeric variables. The dots in a scatter plot not only report the values of individual data points, but also patterns when the data are taken as a whole.**
* **A scatter plot can also be useful for identifying other patterns in data. We can divide data points into groups based on how closely sets of points cluster together. Scatter plots can also show if there are any unexpected gaps in the data and if there are any outlier points. This can be useful if we want to segment the data into different parts, like in the development of user personas.**



3.PCA (Personal Computer Aid)

5. Why is it necessary to investigate data? Is there a discrepancy in how qualitative and quantitative data are explored?

6. What are the various histogram shapes? What exactly are ‘bins'?

7. How do we deal with data outliers?

* **We all have heard of the idiom ‘odd one out which means something unusual in comparison to the others in a group.**
  + **Below are some of the techniques of detecting outliers**
  + **Boxplots**
  + **Z-score**
  + **Inter Quantile Range(IQR)**
* **Deleting the values: You can delete the outliers if you know that the outliers are wrong or if the reason the outlier was created is never going to happen in the future. For example, there is a data set of peoples ages and the usual ages lie between 0 to 90 but there is data entry off the age 150 which is nearly impossible. So, we can safely drop the value that is 150.**
* **Using different analysis methods: You could also use different statistical tests that are not as much impacted by the presence of outliers – for example using median to compare data sets as opposed to mean or use of equivalent nonparametric tests etc.**

8. What are the various central inclination measures? Why does mean vary too much from median in certain data sets?

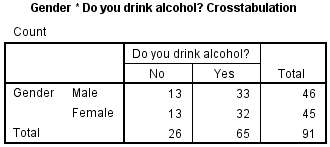
* **Mean, Median, and Mode.**
* **Mean vary too much because when we calculating mean we have to sum all the data and divide with the number of elements if we do so than if there is an outlier in data set than it cause high outcome**

9. Describe how a scatter plot can be used to investigate bivariate relationships. Is it possible to find outliers using a scatter plot?

* **Yes we can find outlier with the help of scatter plot because scatter plot group the similer data in liner regression type .if we have a different data than it will far a way from the all**

10. Describe how cross-tabs can be used to figure out how two variables are related.

* **The dimensions of the crosstab refer to the number of rows and columns in the table. (The "total" row/column are not included.) The table dimensions are reported as as *R*x*C*, where *R* is the number of categories for the row variable, and *C* is the number of categories for the column variable.**
* **Additionally, a "square" crosstab is one in which the row and column variables have the same number of categories. Tables of dimensions 2x2, 3x3, 4x4, etc. are all square crosstabs.**



* **Row variable: *Gender* (2 categories: male, female)**
* **Column variable: *Alcohol* (2 categories: no, yes)**
* **Table dimension: 2x2 (square)**