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

* Collecting Data: As you know, machines initially learn from the data that you give them
* Preparing the Data: After you have your data, you have to prepare it
* Choosing a Model
* Training the Model
* Evaluating the Model
* Parameter Tuning
* Making Predictions.

2. What are the different forms of data used in machine learning? Give a specific example for each of

them.

A:

 numerical data, categorical data, time series data, and text data.

3. Distinguish:

1. Numeric vs. categorical attributes

A:

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.

2. Feature selection vs. dimensionality reduction

A:

While both methods are used for reducing the number of features in a dataset, there is an important difference. Feature selection is simply selecting and excluding given features without changing them. Dimensionality reduction transforms features into a lower dimension.

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

1. The histogram

A:

It provides us a count of the number of observations in each bin created for visualization.

2. Use a scatter plot

A:

A scatter plot is a visual representation of how two variables relate to each other. You can use scatter plots to explore the relationship between two variables, for example by looking for any correlation between them

3.PCA (Personal Computer Aid)

A:

PCA is used to visualize multidimensional data. It is used to reduce the number of dimensions in healthcare data. PCA can help resize an image. It can be used in finance to analyze stock data and forecast returns.

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

data are explored?

A:

This helps them to build accurate models and check assumptions required for fitting models. Create meaningful data visualizations, predict future trends from the data.

The combination of qualitative and quantitative data can also lead to clashes in the philosophical assumptions behind each approach and therefore recommendations have been made for triangulation to be carried out from a pragmatic, or subtle realist, approach.

6. What are the various histogram shapes? What exactly are ‘bins&#39;?

A:

A histogram is a chart that plots the distribution of a numeric variable's values as a series of bars. Each bar typically covers a range of numeric values called a bin or class; a bar's height indicates the frequency of data points with a value within the corresponding bin.

7. How do we deal with data outliers?

A:

* **Dropping the outliers** - it prevents skewing of the data
* **Cap them** - ie. define a max/min point and assign that value to the outlier. This holds if the data suggests that after some point the bigger/smaller value did not change the outcome (ie. should you buy bubblegum if your salary is 1M or 10M?)
* **Double check them** - they could be wrong. In that case, you may input as missing/mean/median value.
* **Change the scale** - use normalization, etc.

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

certain data sets?

A:

* mode.
* median.
* mean.

Mean is impacted by outliers , also it varies widely depending on the data set available in the sample set.

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

outliers using a scatter plot?

A:

A large amount of scatter around the line indicates a weak relationship. Little scatter represents a strong relationship. If all points fall directly on a straight line, we have a perfect linear relationship between our two variables.

Use this chart when there are many different data points, and you want to highlight similarities in the data set. This is useful when looking for outliers or for understanding the distribution of your data.

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

A:

Cross-tab maps out the correlation between variables, insights that otherwise may have been overlooked are clearly understood. It is straightforward to understand the insights from even a complicated form of statistics. It provides qualified or relative data on two or more variables across multiple features with ease.