1. **Source**:
   We are getting our data from a kaggle dataset. It represents a collection of used car transactions with details about the car. The purpose of collecting this data is to build a database and predictive model that users can interact with to get an accurate price for a used car. The data is publicly available and there are no restrictions on its use that we have found. DATA

2. **Structure and Metadata**:
   The key features of our dataset are condition, odometer, make, model and selling price. These columns are the ones we predict will be most useful in our prediction process. The dataset is 550298 records in total and some key metrics about our data and columns are present in our EDA notebook which will be submitted as well.

3. **Missing Data**:
   - Are there any missing values? If so, how are they represented?
   - What strategies can be used to handle missing data?

   Here is a summary of missing values in our dataset. Some columns are missing a significant number of values. This is an issue we plan to address in multiple different ways. For numerical columns we are planning on using the mean of that column to fill in missing values. For certain categorical columns such make/model/trim we may have to drop the entire index since it will be very difficult to infer these based on other columns. For other columns like transmission we can look at other records present in the dataset to populate missing value. We want to be in a sweet spot between dropping rows vs assuming/calculating missing data based on available related data. Dropping too many rows will not be ideal as it can limit data for predictive modeling and assuming and filling in too many values can also affect us as we might end up having more inaccurate representation which will further affect prediction. Something to note here are the seller and the state column as they might not really add insight into model prediction or getting an accurate price for the car; whereas the year column being empty we are considering to calculate based on related columns we might calculate/assume them as they are one of the key factors to help us understand and determine the price of the used car.

4. **Anomalies**:
    - Are there outliers or anomalies? do they indicate errors?
    - How consistent is the data across features and observations?

    [HAVE WE DETECTED ANOMALIES? ARE THEY RELEVANT?]

5. **Bias**:

    This dataset is going to be biased towards more popular car brands and models. It is going to perform better on cars that sold more units since there is more data. This discriminates against higher level/luxury vehicles that may only sell a significantly less number of cars per year to a smaller client base. We are choosing to ignore this since we want our work to be used by regular people so we feel that we are not losing too much functionality in our project by a lower performance on higher end vehicles.

6. **Distributions**:

    Some columns in our data are going to be correlated such as odometer and year or mmr (car rating) and price. These correlations are inevitable and part of the domain. Our numerical features have significant skews that you would expect with our data. Odometer reading is skewed right since most cars with lots of miles are probably not being sold. The mmr and price are skewed similar to odometer with most values being on the left and right tail.

7. **Categorical Data**:

    Each categorical column contains a large number of categories due to how many different options there are for cars. Columns like make and color have many relevant categories but other columns such as trim have many entries with only 1 trim level. Most categorical columns are not well balanced.

8. **Ethical Considerations**:

    There are no ethical concerns in our data and we should have no problem publishing our findings and insights. There is no sensitive or personally identifiable information in our dataset.

10. **Alignment with Goals**:

This dataset aligns very well with our goals. It contains a lot of relevant features about car sales and should allow us to create effective models.

The dataset contains necessary features for the intended analysis. We will discuss more on which models will be used and which features are most useful, later.

11. **Scalability**:

    The dataset is manageable with our available resources. With only about 500000 records we have not had any trouble loading and working with the data. We may run into an issue if we choose to apply deep learning to our ensemble method but we have access to the discovery cluster if we need it for more compute power.

12. **Transformations**:
    - Does the data need preprocessing, such as normalization, standardization, or scaling?
    - Are there opportunities to create new features that could improve the model or analysis?

    [ADD TRANSFORMATIONS TO DATA HERE ONCE WE FINALIZE]
13. **Data Encoding**:
    - How should categorical variables be encoded (e.g., one-hot, label encoding)?
    - Are there any temporal or sequential features that require specific transformations?
14. **Predictive Power**:
    - Do the features contain sufficient predictive information for the target variable?
    - Is feature selection or dimensionality reduction necessary?
15. **Target Variable**:
    - If this is a supervised learning problem, what is the target variable, and is it well-defined?
    - Is the target variable balanced, or does it require special handling (e.g., resampling)?
16. **Validation Strategy**:
    - How will you split the dataset for training, validation, and testing?

- Are there temporal or spatial dependencies that need to be preserved during splitting?

We are going to split our dataset into training testing and validation using the sklearn train test split function. This should allow us to easily partition our data to test our models. We are not aware of any temporal or spatial dependencies that need to be preserved.

17. **Data Leakage**:
    - Are there any risks of data leakage, where information from the test set inadvertently influences the model during training?

18. **Interpretability**:
    - Can the dataset and analysis provide interpretable insights for stakeholders?
    - How will the results be communicated (e.g., visualizations, metrics)?

The results will be communicated through the dashboard in the form of visualizations and metrics.

19. **Limitations**:
    - What are the limitations of this dataset for the current project?
    - What additional data would enhance the analysis?

A major limitation of our current project is that we can only find data up till 2015 which will make it hard to predict pricing of cars built after 2015. We are exploring possible data transformations to try and combat this and actively looking for more data to try and counteract this problem.

Similarly formatted dataset with more recent years will be much more useful to enhance the analysis.