ipynb/01-Ingest Data

1. Find five features in the dataset description that you think will be important to predicting the sale price of a home. Describe why you think each of these features is important.

2. Describe how the `merge` function works.

3. What was done with the `Id` feature?

4. Why did so many features need to be converted to factors?

5. Are there other features that you feel should be converted to a factor? Why?

6. Make the change for any features that you feel should be factors. Make sure to change this notebook, `src/load_data-01.r` and `src/load_data-02.r`.

ipynb/02-impute_nan_values.ipynb

1. What is done by `source('../src/load_data-01.r')`?

2. What does this do?

```
nan_sums = colSums(is.na(housing_df))
nan_sums[nan_sums > 0]
```

3. What does this do?

```
mean_LotFrontage <- mean(housing_df$LotFrontage, na.rm=T)
mean_MasVnrArea <- mean(housing_df$MasVnrArea, na.rm=T)
mean_GarageYrBlt <- mean(housing_df$GarageYrBlt, na.rm=T)

housing_df$LotFrontage[is.na(housing_df$LotFrontage)] <- mean_LotFrontage
housing_df$MasVnrArea[is.na(housing_df$MasVnrArea)] <- mean_MasVnrArea
housing_df$GarageYrBlt[is.na(housing_df$GarageYrBlt)] <- mean_GarageYrBlt
```

4. Is there a better strategy?

5. What does this do?

```
count_empty_values <- function (feature) {
    empty_string_mask = housing_df[feature] == ""
    return(length(housing_df[feature][empty_string_mask]))
}
```

6. Why is this necessary?

7. What is this?

```
empty_means_without <-c("Alley","BsmtQual","BsmtCond","BsmtExposure","BsmtFinType1",
                        "BsmtFinType2", "FireplaceQu","GarageType","GarageFinish",
                        "GarageQual","GarageCond","PoolQC","Fence","MiscFeature")

empty_means_NA <- c("MasVnrType","Electrical")
```

8. What does this do?

```
housing_df <- na.omit(housing_df)
```

ipynb/03-basic_eda.ipynb

1. Why did I run this?

```
count_empty_total()
```

2. Plot a histogram with a KDE Plot for as many numerical features as you think are interesting or important.

3. Plot histograms for Sale Price sorted by category for as many categorical features are you think are interesting or important.

4. Find five features in the dataset description that you think will be important to predicting the sale price of a home. Describe why you think each of these features is important.

5. Compare the five that you identified during plotting to the five you discussed at the beginning. How does your intuition compare to what you see in the Distribution plots.