Goals of the Week 2 Project

- Learn & understand
  - Models (logistic regression, trees & forests)
  - Model evaluation (confusion matrix , model performance metrics)
  - Cross-validation
  - Exploratory analyses
  - Feature engineering (imputation, dummy coding/OHE, binning, interaction)
    - preferably with sklearn
    - What I did ?
      - imputation, dummy coding with pandas
      - interaction with sklearn
  - Pipeline usage
    - with sklearn
- Participate in the Kaggle challenge

My manual pipeline for the test.csv

- Do the data exploration on the whole test data (test.csv) -- "data snooping", "data torturing"
- Designate a few potential candidate features for prediction
  - Do the imputation, dummy coding with pandas on the whole test data (test.csv)
- Apply random forest to the whole test data (test.csv) for empirical feature selection
- Apply the sklearn pipeline
  - test, train split on (test.csv)
  - interaction
  - cross-validation
- See the performance of different models
  - Of course random forest beats logistic regression

Importance of Data Exploration

- Advantages & disadvantages
  - Get a good sense of your data
    - Doable with small data-frames like Titanic
    - Not very doable with big data
  - (Borderline)-cheating
    - potential overfitting / non-generalizability issues

Problems I experienced

Creating the pipeline in Sklearn due to

… went away after updating to JupyterLab to v. 3.0.5

… before that it was all sklearn errors during the FE

- (probably) not somehting to do with NaN handling first and then doing the rest

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 891 entries, 1 to 891
Data columns (total 11 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Survived  891 non-null    int64
 1   Pclass    891 non-null    int64
 2   Name      891 non-null    object
 3   Sex       891 non-null    object
 4   Age       714 non-null    float64
 5   SibSp     891 non-null    int64
 6   Parch     891 non-null    int64
 7   Ticket    891 non-null    object
 8   Fare      891 non-null    float64
 9   Cabin     204 non-null    object
 10  Embarked  889 non-null    object
dtypes: float64(2), int64(4), object(5)
memory usage: 83.5+ KB
```

```
# df = df.convert_dtypes()
# df.info()
```

The accuracy of the CV RF model for the test data (test part of train.csv) is 0.87, which I'm skeptic about...

... will see what Kaggle has to say; I'll need to update my notebook for submission

Goals of the Week 2 Project

- Learn & understand
  - Models (logistic regression, trees & forests)
  - Model evaluation (confusion matrix , model performance metrics)
  - Cross-validation
  - Exploratory analyses
  - Feature engineering (imputation, dummy coding/OHE, binning, interaction)
    - preferably with sklearn
    - What I did ?
      - imputation, dummy coding with pandas
      - interaction with sklearn
  - Pipeline usage
    - with sklearn
- Participate in the Kaggle challenge