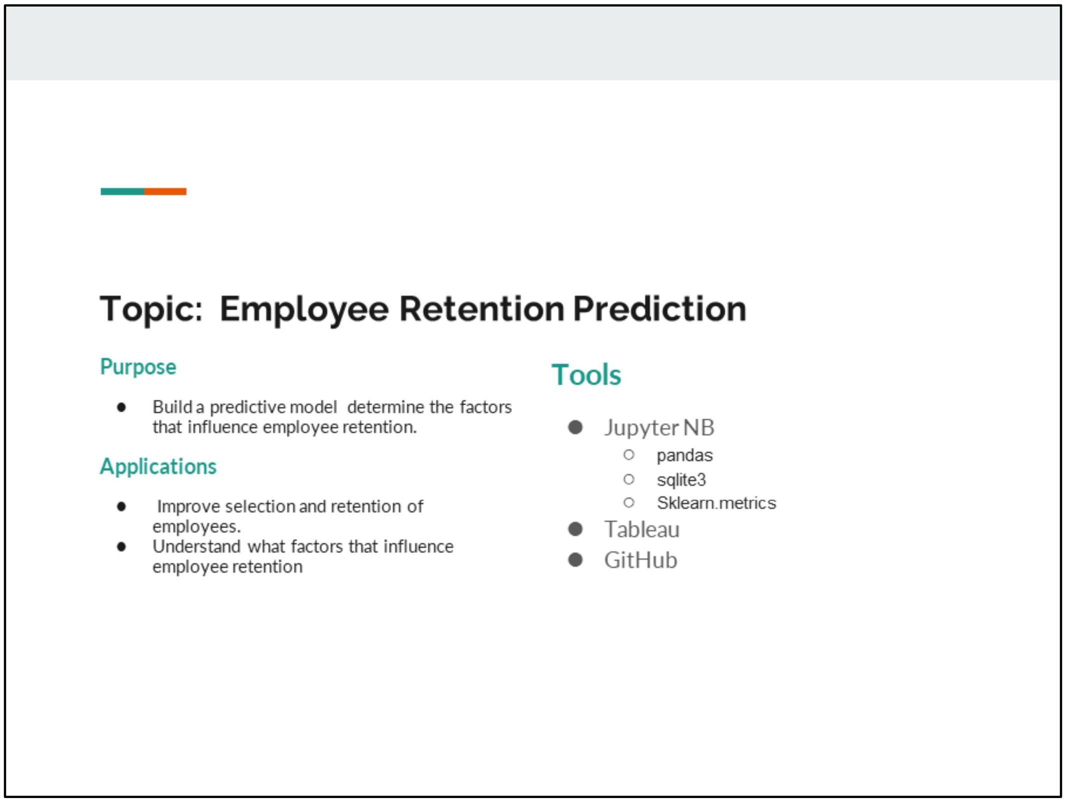
Factors Influencing Employee Retention

Griselda Arevalo

Coral Rosa-Falero

Jiemin Sheng

Bootcamp: GWU-VIRT-DATA-PT-08-2022-U-B-MW

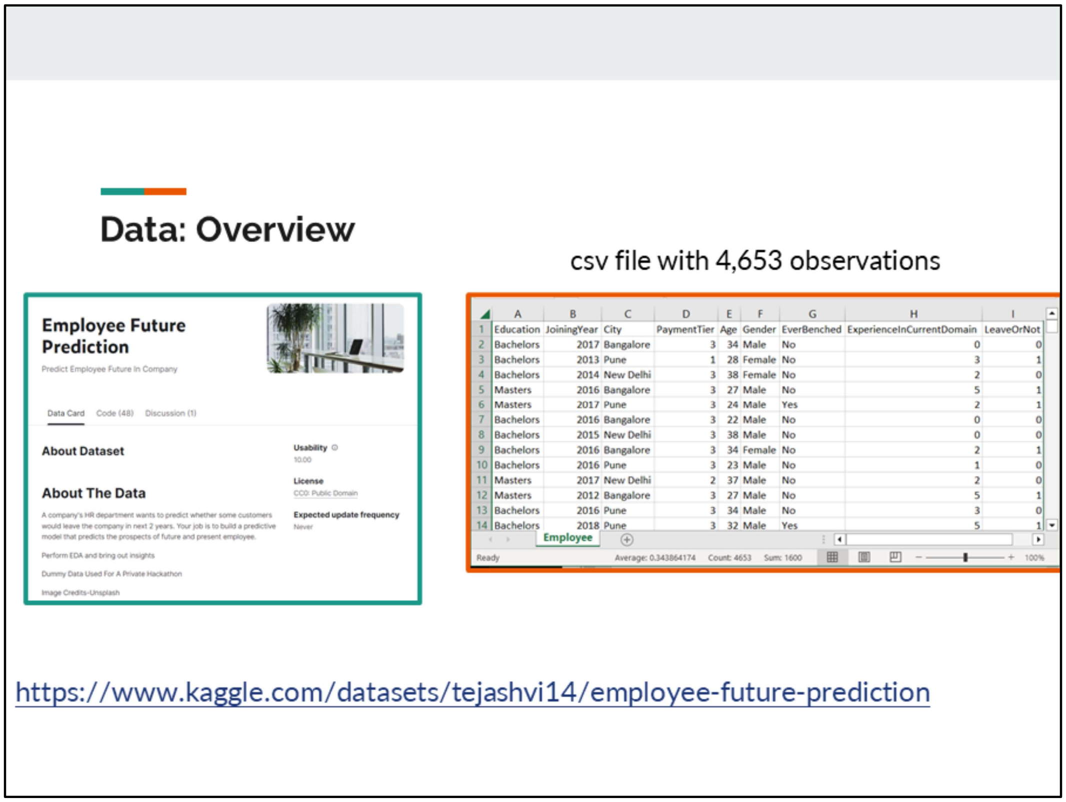


Purpose

- Build a predictive model determine the factors that influence employee retention.

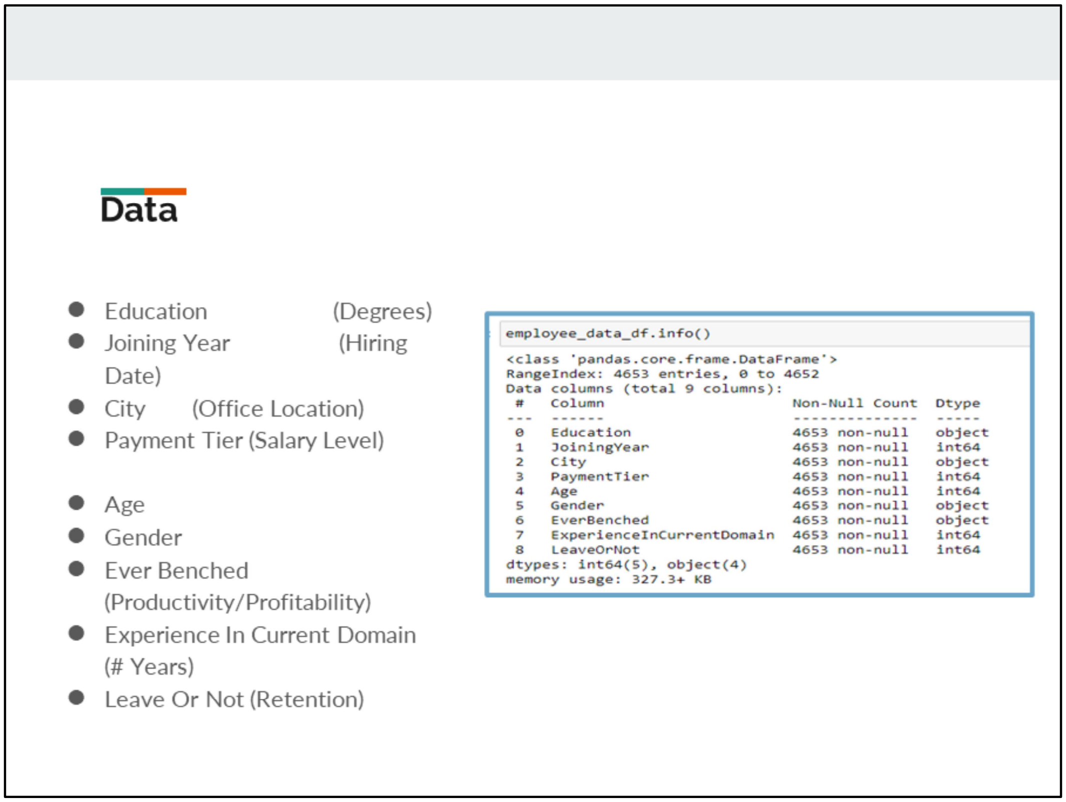
Applications

- Assist human resource departments improve selection of future employees and improve the retention of new employees.
- To better understand what variables influence the employees either stay or leave.

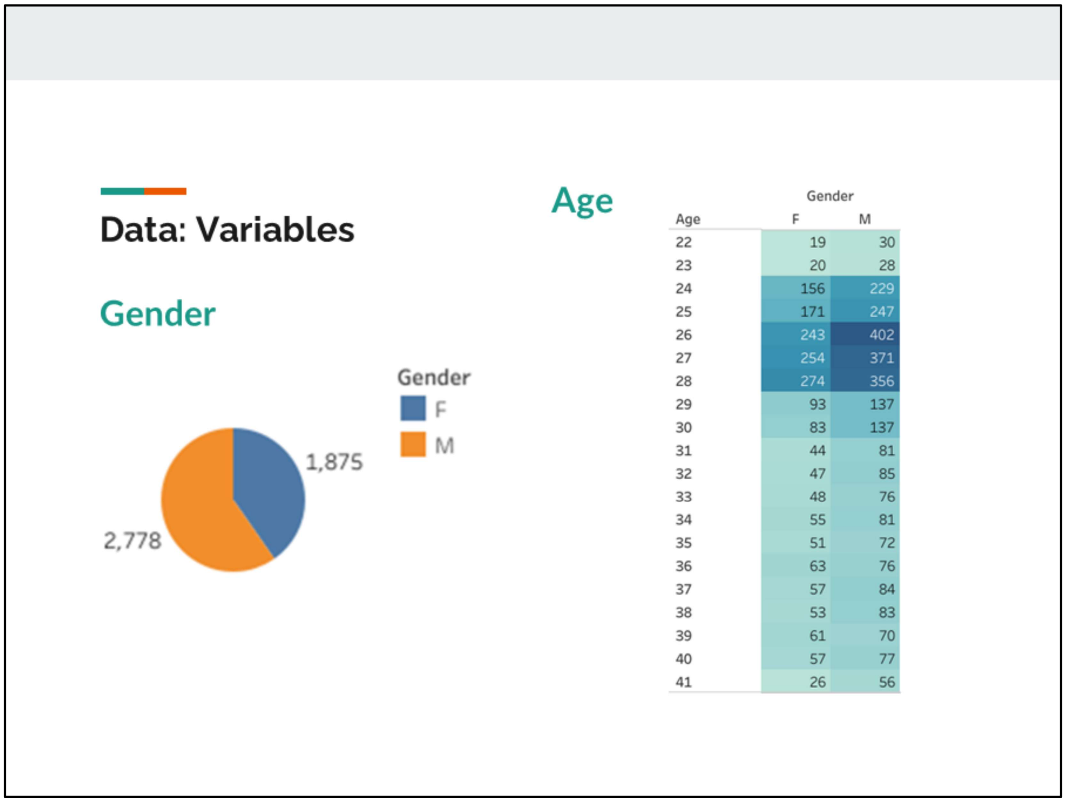


1. Clean csv file from kaggle
2. Selected because high usability score, database completeness and cleanliness

The firm in question does not appear to be stated but can be noted to be located in India based on the office locations in Indian cities.

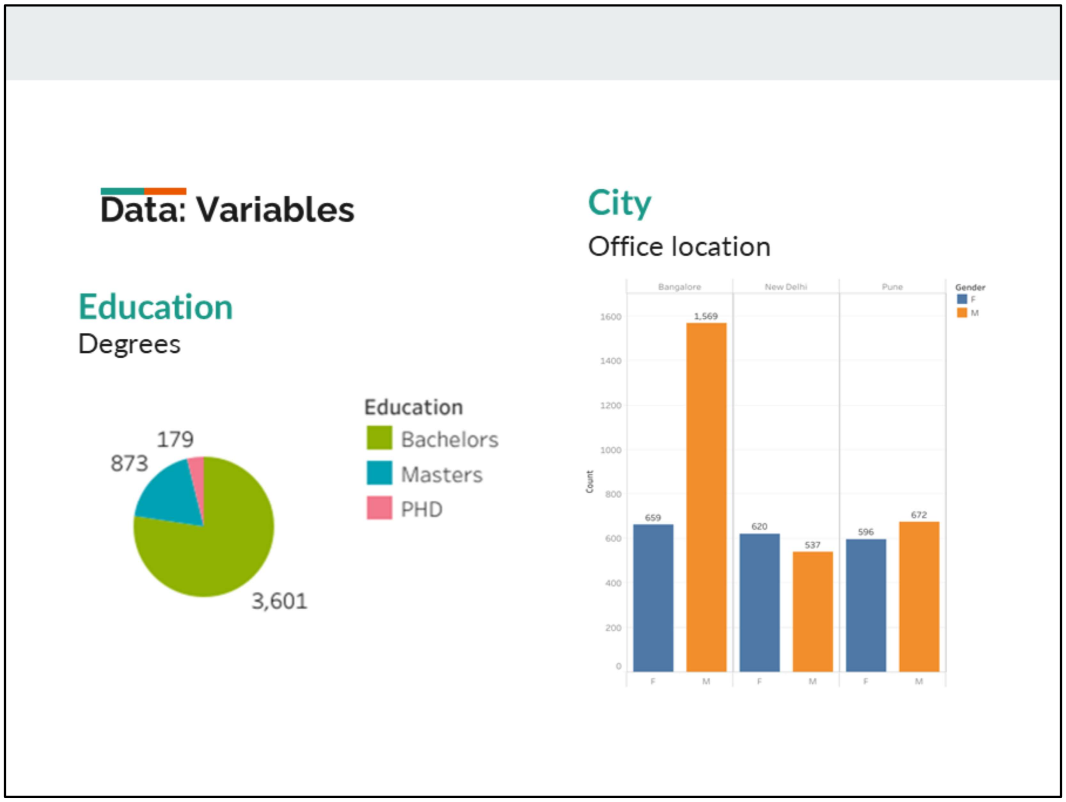


1. There is a total of 4,653 observations.
2. Education level is a qualitative categorical variable consisting of three possible values: Bachelors, Masters, and Doctorate degrees.
3. The year of joining firm variable is a continuous quantitative variable varies from 2012 to 2018.
4. The city location of office is a qualitative categorical variable consisting of three cities: Bangalore, Pune, and New Delhi.
5. The salary tier is a qualitative categorical variable which include three tiers: 1 for the highest, 2 for the middle, and 3 for the lowest. Age is a continuous quantitative variable ranging from 22 to 41.
6. The gender variable is qualitative and categorical with either male or female.
7. The qualitative categorical variable of whether or not the employee was kept out of projects for longer than one month is either yes or no.
8. The experience length variable is continuous and quantitative and ranges from 0 to 7.
9. The dependent variable being predicted of whether or not the employee is expected to be leaving the firm within two years is qualitative categorical variable of either 1 or 0 for yes or no.categorical variable of either 1 or 0 for yes or no.



Age	Gender F	Gender M
22	19	30
23	20	28
24	156	229
25	171	247
26	243	402
27	254	371
28	274	356
29	93	137
30	83	137
31	44	81
32	47	85
33	48	76
34	55	81
35	51	72
36	63	76
37	57	84
38	53	83
39	61	70
40	57	77
41	26	56

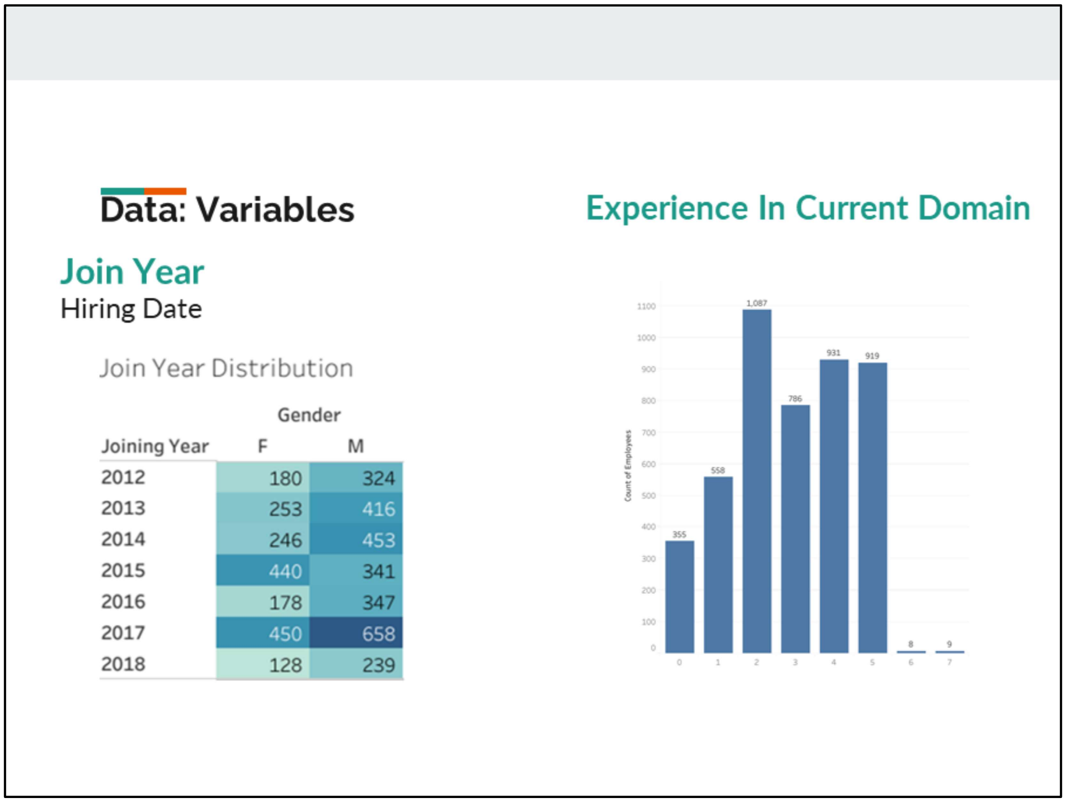
1. The **gender** variable is qualitative and categorical with either male or female
   a. More Males than Females
2. **Age** is a continuous quantitative variable ranging from 22 to 41.
   a. Most of the employees are within the ages of 24 and 28



1. **Education** level is a qualitative categorical variable consisting of three possible values: Bachelors, Masters, and Doctorate degrees
    a. Great majority posses a BS degree
2. The **city** location of office is a qualitative categorical variable consisting of three cities: Bangalore, Pune, and New Delhi.
    a. The bangalore office contains the majority of the employees.
    b. New Delhi and Pune with comparable numbers of M/F. while in the Bangalore office the M employees are vet twice as many as F employees

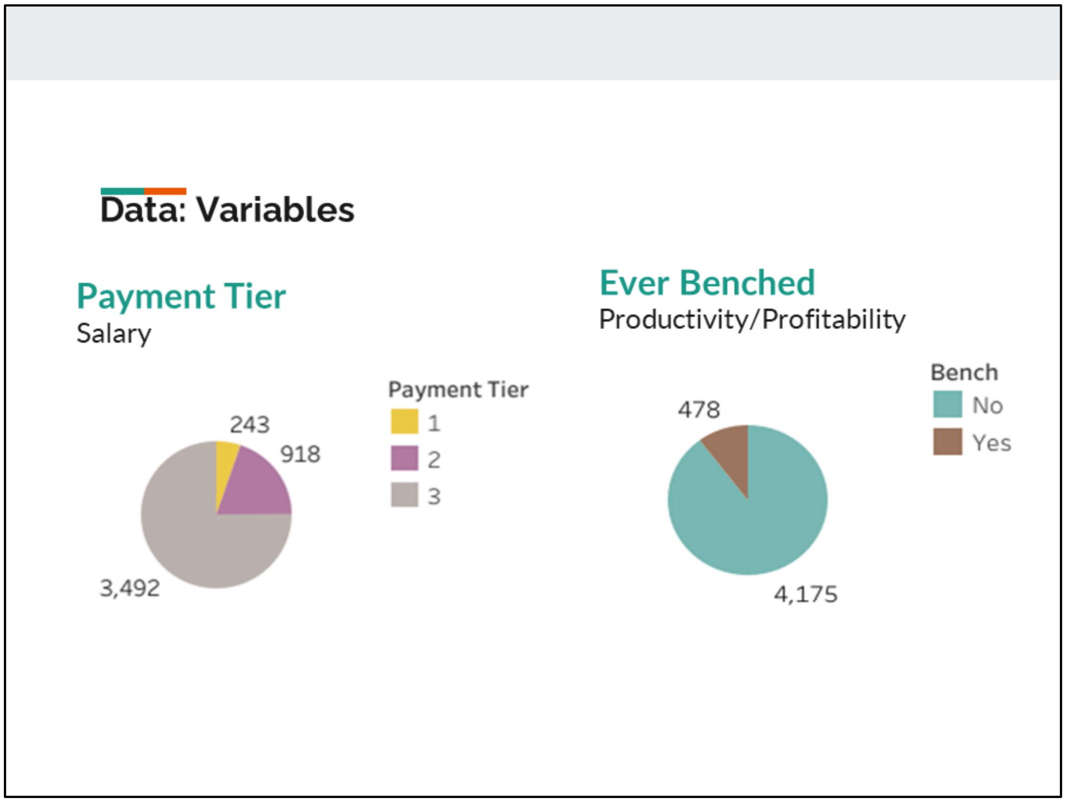
===

Both images are set up to be filtered. Image can be replaced with link to tableau for live demo

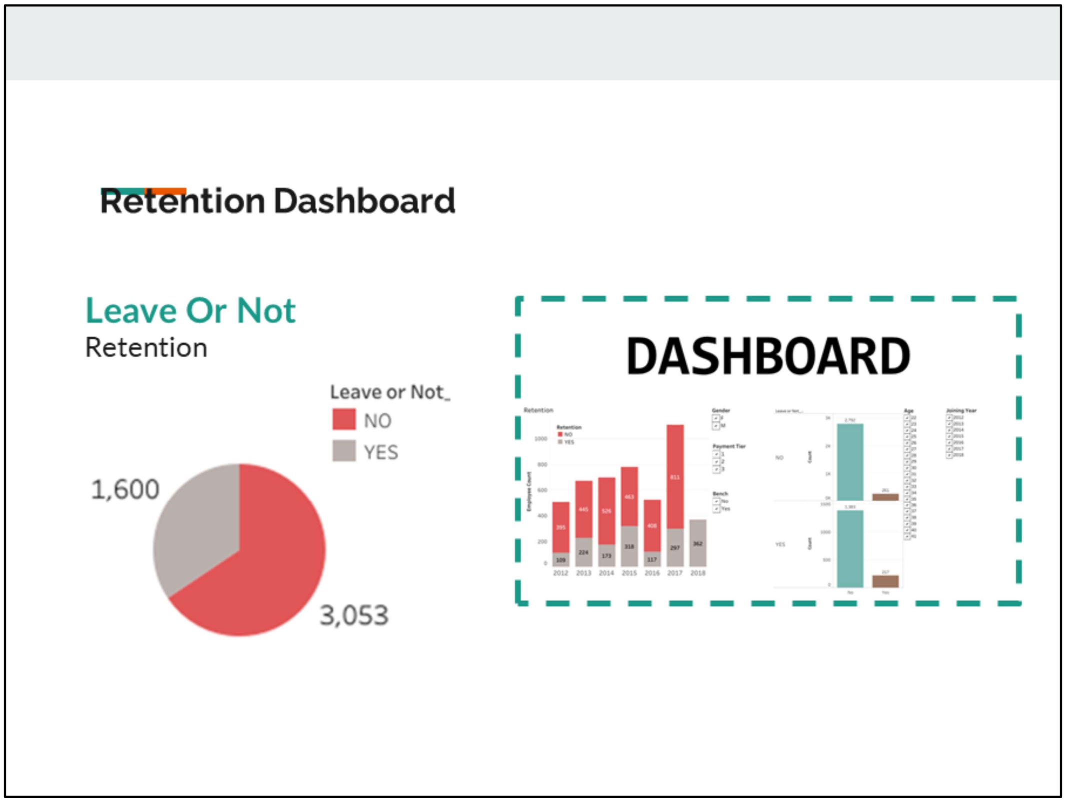


Joining Year	F	M
2012	180	324
2013	253	416
2014	246	453
2015	440	341
2016	178	347
2017	450	658
2018	128	239

1. The **year of joining** firm variable is a continuous quantitative variable varies from 2012 to 2018.
    a. Hiring increased from 2012 to 2015, dipped in 2016, doubled in 2017 and dropped again during 2018
2. The **experience** length variable is continuous and quantitative and ranges from 0 to 7y
    a. The majority of employees have up to 5 years of experience in the current field.
    b. A disproportionate minority has 5+ years experience
    c. Most employees fall in the 2, 4 and 5y bracket and this distribution is consistent between genders



1. The **salary** tier is a qualitative categorical variable which include three tiers: 1 for the highest, 2 for the middle, and 3 for the lowest
    a. Lowest tier contains more employees than the other 2 tiers combined
2. The qualitative categorical variable of whether or not the employee was kept out of projects (**ever benched**) for longer than one month is either yes or no.
    a. Benching similar between M/F
    b. Highest benching was in 2015



1. The dependent variable being predicted of whether or not the employee is expected to be leaving the firm within two years (**retention**) is qualitative categorical variable of either 1 or 0 for yes or no
   a. ???????

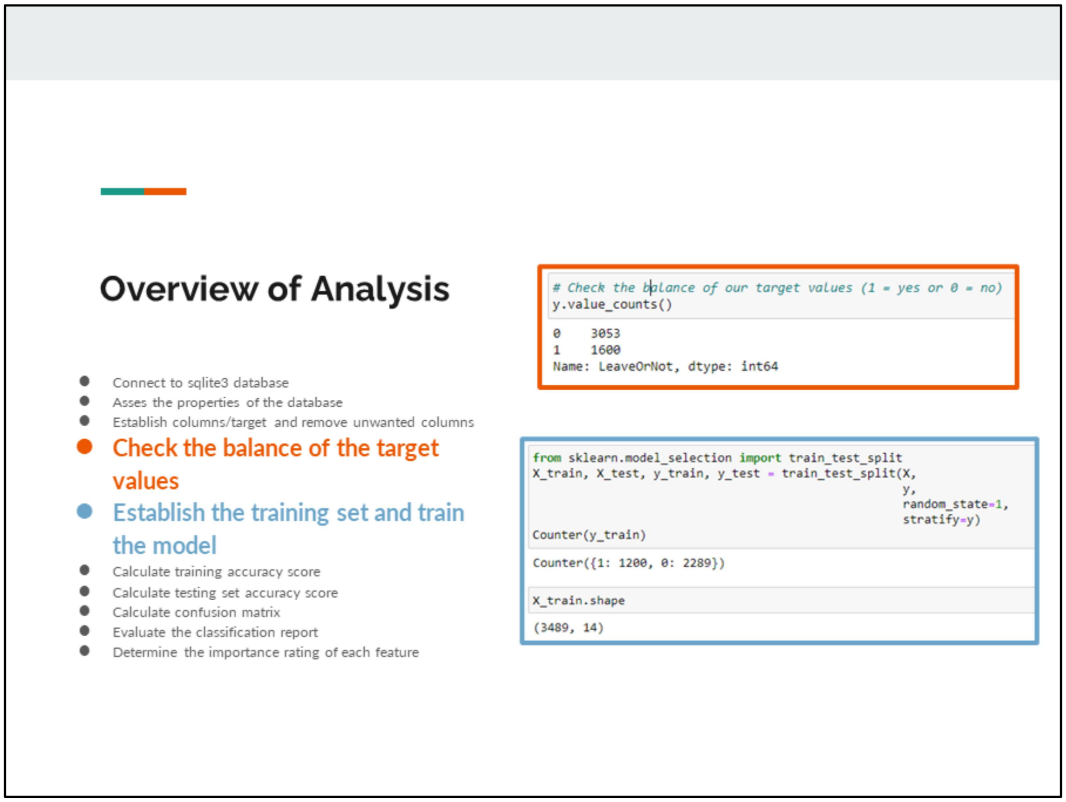
Click on dashboard to open the interactive view. Use the filters and discuss

Overview of Analysis

- Connect to sqlite3 database
- Asses the properties of the database
- Establish columns/target and remove unwanted columns
- Check the balance of the target values
- Establish the training set and train the model
- Resample the training data
- Calculate training accuracy score
- Calculate testing set accuracy score
- Calculate confusion matrix
- Evaluate the classification report
- Determine the importance rating of each feature

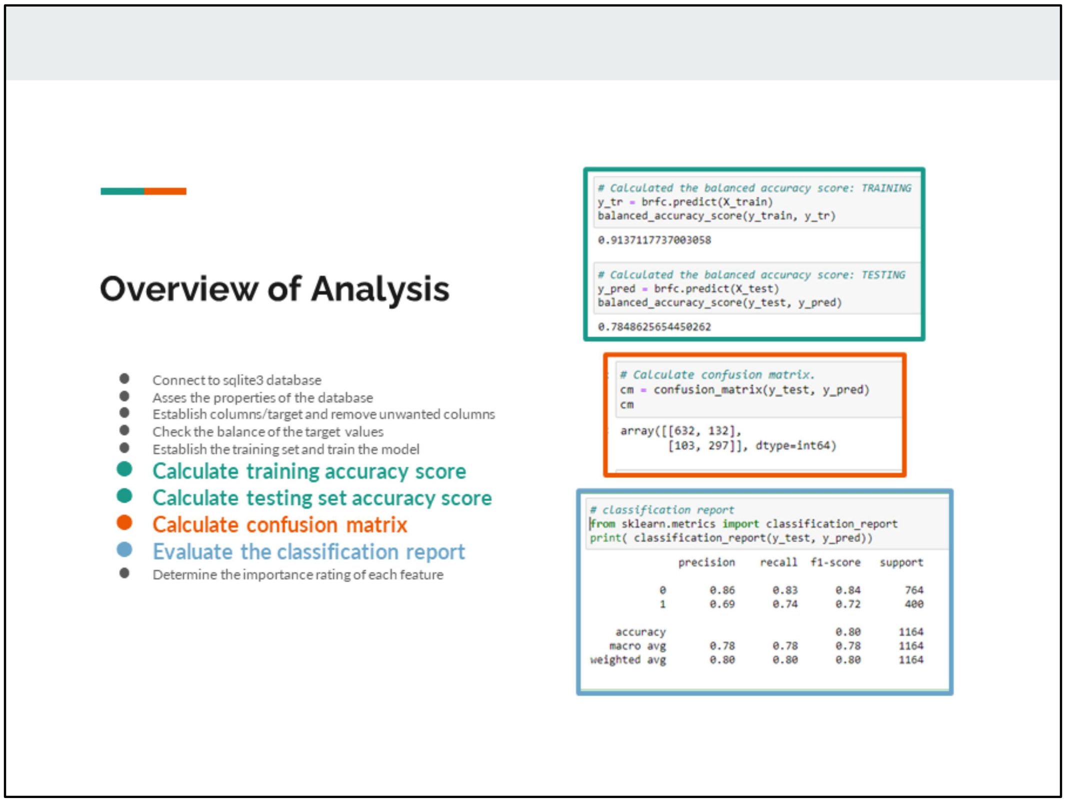
```
employee_data_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4653 entries, 0 to 4652
Data columns (total 9 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Education                  4653 non-null   object
 1   JoiningYear                4653 non-null   int64
 2   City                       4653 non-null   object
 3   PaymentTier                4653 non-null   int64
 4   Age                        4653 non-null   int64
 5   Gender                     4653 non-null   object
 6   EverBenched                4653 non-null   object
 7   ExperienceInCurrentDomain  4653 non-null   int64
 8   LeaveOrNot                 4653 non-null   int64
dtypes: int64(5), object(4)
memory usage: 327.3+ KB
```



The observations for target 0 are 2x as those for target 1 therefore seasonably balanced but more experience is required to properly determine if this is a correct assumption

The train/test split is 75/25 and the y_train maintains the same pattern of 2 to 1 for 0 and 1.
Tain set (3489, 14), test set (1164, 14) [3489 rows of data with 14 variables]



Accuracy score: how many predictions did the model got correct. *For binary classifications, balanced accuracy is equal to the arithmetic mean of sensitivity (true positive rate) and specificity (true negative rate)**

Accuracy score training: 91%,
Accuracy score test : 78%
====================================
PRECISION: How reliable a positive classification iis : Positive Predictive Value (PPV) for 0 = 86% and 1 = 69%

SENSITIVITY: How many items were correctly evaluated: 0 = 83% and 1 = 74%

SPECIFICITY: (true negative rate/ *recall of the negative class***

Overall, the model is better at predicting 0 than 1

F1 score combines precision and sensitivity, low value reflects major differences between precision and sensitivity. 80

Classification_report and classification_reporrt imbalance yielded similar results; unclear of the implications of that.

References

*https://scikit-learn.org/stable/modules/model_evaluation.html#classification-report

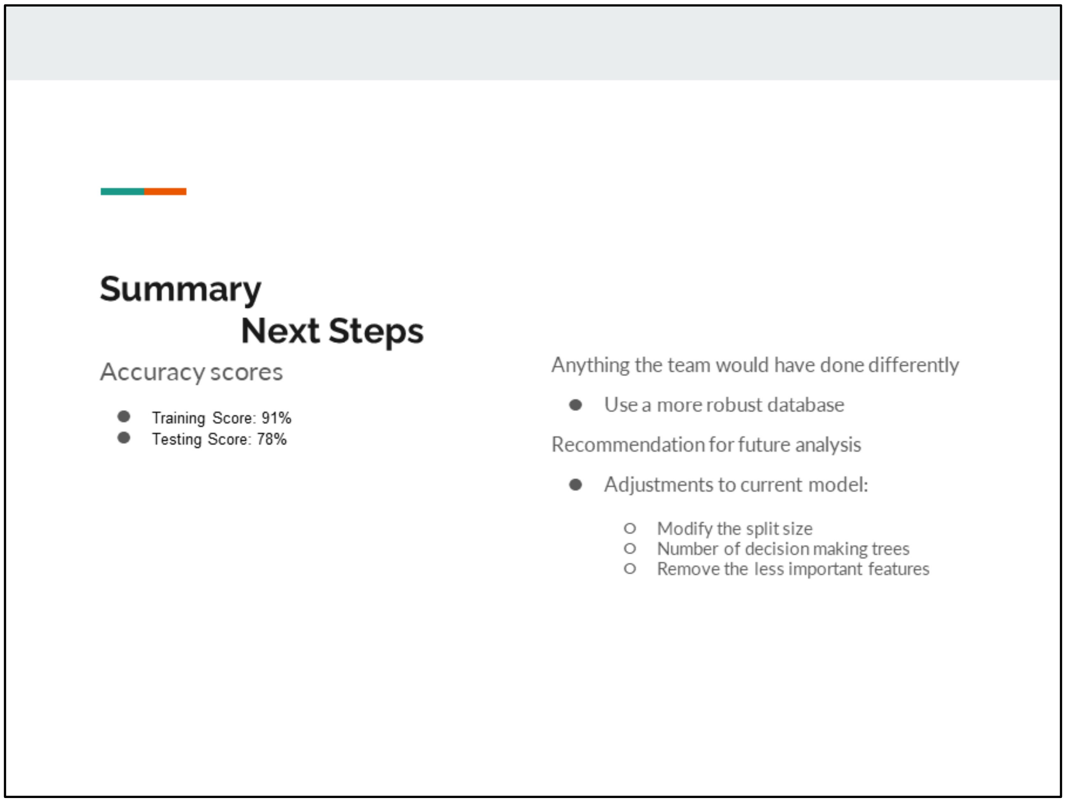
**https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html
Contains additional calculations that can be explored as the future analysis

Overview of Analysis

- Connect to sqlite3 database
- Asses the properties of the database
- Establish columns/target and remove unwanted columns
- Check the balance of the target values
- Establish the training set and train the model
- Calculate training accuracy score
- Calculate testing set accuracy score
- Calculate confusion matrix
- Evaluate the classification report
- **Determine the importance rating of each feature**

```python
# List the features sorted in descending order by feature importance
importances = brfc.feature_importances_
sorted(zip(brfc.feature_importances_, X.columns), reverse=True)
```

```
[(0.3086914210245113, 'JoiningYear'),
 (0.19190147118381745, 'Age'),
 (0.09890156348096112, 'ExperienceInCurrentDomain'),
 (0.09325636092904764, 'PaymentTier'),
 (0.05727410779116132, 'City_Pune'),
 (0.054031392148976856, 'Education_Masters'),
 (0.046093613205681915, 'Gender_Male'),
 (0.03876883320695893, 'Education_Bachelors'),
 (0.03204738082388843, 'Gender_Female'),
 (0.025990119766903574, 'City_Bangalore'),
 (0.025687810889074857, 'City_New Delhi'),
 (0.009464044355872172, 'EverBenched_No'),
 (0.009311635638471467, 'EverBenched_Yes'),
 (0.008580245554672804, 'Education_PHD')]
```



Anything the team would have done differently

Use a more robust database with sufficient background information.

This database although clean and complete (no null values) seems to only focus on a very limited set of factors that seem more sided with what an Employer would consider when deciding to retain an employee and does not evaluate factors the Employees might be more included to consider when deciding to remain at a certain job including (1)training and development opportunities, (2) flexible telecommuting, (3) company culture/environment, (4) administrative/restructuring /leadership changes amongst others.

Some of the data is difficult to interpret due to lack of context, for example:

(1) the office location might be more significant if the database included residence information for each employee

(2) the salary tier is not informative because there are no details about the actual salary contained within the tiers