Data Analytics in Python and SQL: Final Project

Guillermo Casillas, Valeria Chavez, Benjamin Catton

Background Information

1

Two Major Datasets: IPO and Funding Rounds

2

Impact of Investor Behavior and Market Trends

3

Compare the path from startup to IPO

4

VC Firms, startup founder, PE, and analysts

Set-Up

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1259 entries, 0 to 1258
Data columns (total 13 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   id                       1259 non-null   int64
 1   ipo_id                   1259 non-null   int64
 2   object_id                1254 non-null   object
 3   valuation_amount         1259 non-null   float64
 4   valuation_currency_code  1257 non-null   object
 5   raised_amount            1259 non-null   float64
 6   raised_currency_code     699 non-null    object
 7   public_at                659 non-null    object
 8   stock_symbol             1259 non-null   object
 9   source_url               191 non-null    object
 10  source_description       180 non-null    object
 11  created_at               1259 non-null   object
 12  updated_at               1259 non-null   object
dtypes: float64(2), int64(2), object(9)
memory usage: 128.0+ KB
```

```
[5]: fr_df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 52928 entries, 0 to 52927
Data columns (total 23 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   id                       52928 non-null  int64
 1   funding_round_id         52928 non-null  int64
 2   object_id                52928 non-null  object
 3   funded_at                52680 non-null  object
 4   funding_round_type       52928 non-null  object
 5   funding_round_code       52928 non-null  object
 6   raised_amount_usd        52928 non-null  float64
 7   raised_amount            52928 non-null  float64
 8   raised_currency_code     49862 non-null  object
 9   pre_money_valuation_usd  52928 non-null  float64
 10  pre_money_valuation      52928 non-null  float64
 11  pre_money_currency_code  26883 non-null  object
 12  post_money_valuation_usd 52928 non-null  float64
 13  post_money_valuation     52928 non-null  float64
 14  post_money_currency_code 30448 non-null  object
 15  participants             52928 non-null  int64
 16  is_first_round           52928 non-null  int64
 17  is_last_round            52928 non-null  int64
 18  source_url               40382 non-null  object
 19  source_description       43439 non-null  object
 20  created_by               48291 non-null  object
 21  created_at               52928 non-null  object
 22  updated_at               52928 non-null  object
dtypes: float64(6), int64(5), object(12)
memory usage: 9.3+ MB
```

Summary

- Set-Up pandas and imported our CSV files
- After we ran a quick .info() to see the variables we were working with and their types

Data Analysis

Average Stats by IPO Outcome

went_ipo_max	raised_amount_usd_sum	participants_sum
Did not IPO	11,581,749	3
Went IPO	108,449,318	4

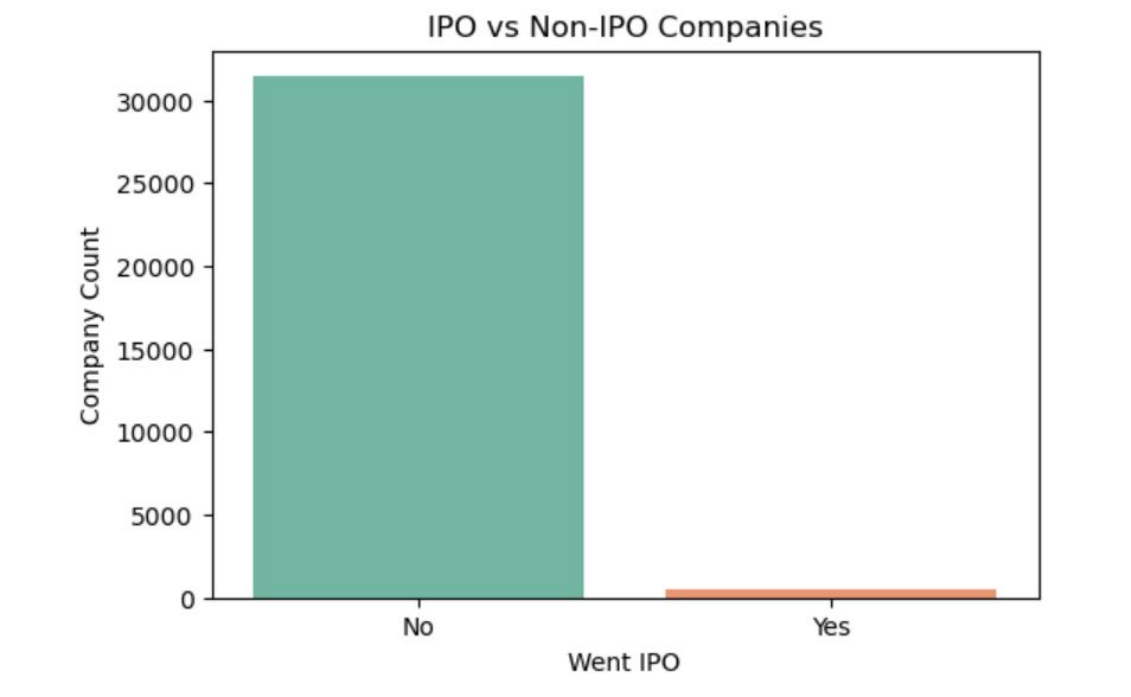
Top 10 Companies by Total Raised

	object_id	raised_amount_usd_sum
987	c:13219	$5,700,000,000
24634	c:4843	$3,985,050,000
10179	c:216492	$3,822,518,000
13375	c:242735	$2,600,000,000
25053	c:5	$2,425,700,000
27947	c:64365	$2,400,000,000
11198	c:22568	$1,765,504,319
27048	c:5951	$1,451,000,000
21775	c:39799	$1,270,283,000
13911	c:24693	$1,200,000,000

Summary

- Average Stats by IPO Outcome
- Top 10 Companies by Total Raised

Data Analysis



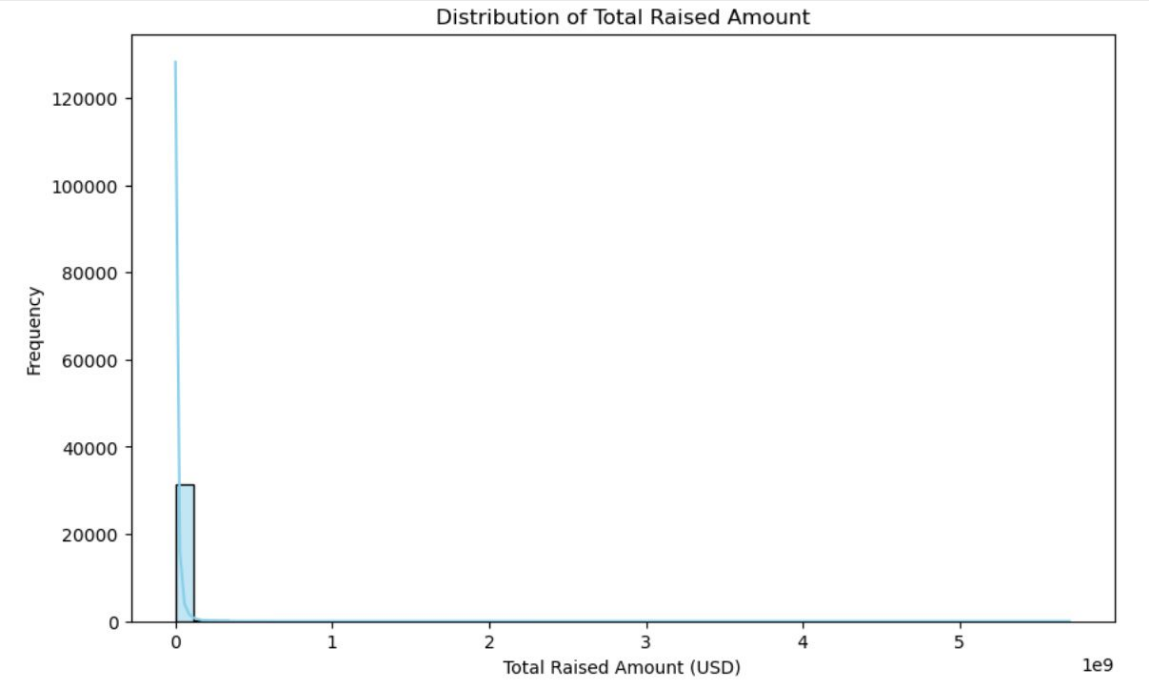
Data Analysis

	Number of Rounds	Number of Companies
Distribution of Funding Rounds per Company		
0	1	20,721
1	2	6,171
2	3	2,671
3	4	1,220
4	5	603
5	6	249
6	7	147
7	8	67
8	9	46
9	10	23
10	11	10
11	12	3
12	13	5
13	14	1
14	15	2

Data Analysis

funding_round_type	avg_raised_usd_million
angel	0.31
crowdfunding	1.64
other	11.24
post-ipo	169.4
private-equity	25.02
series-a	5.91
series-b	11.34
series-c+	21.17
venture	8.16

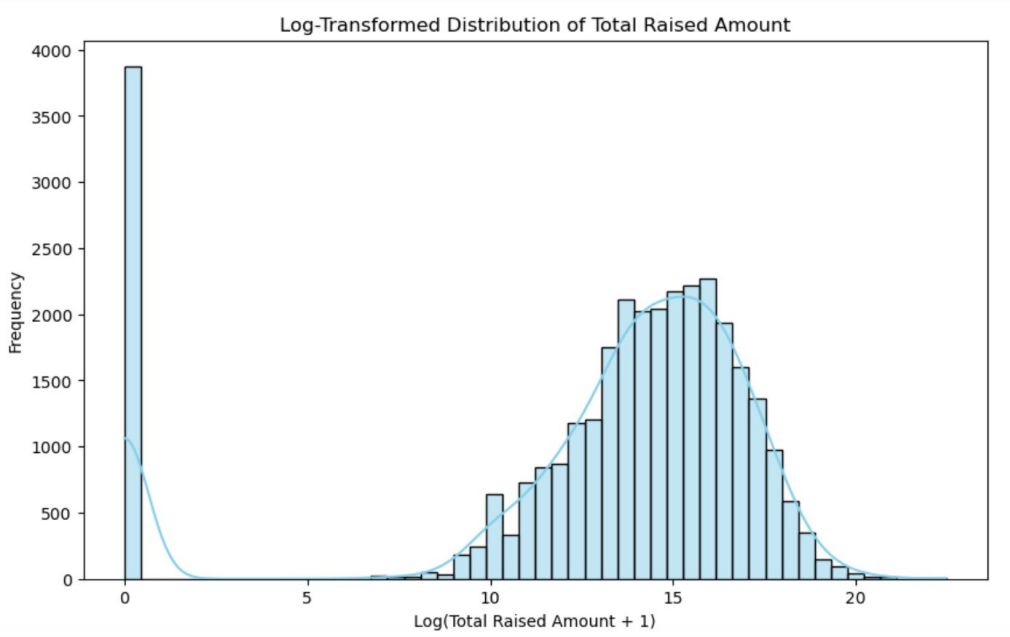
Data Analysis



Summary

- Doesn’t tell us much so we need to use a transformation

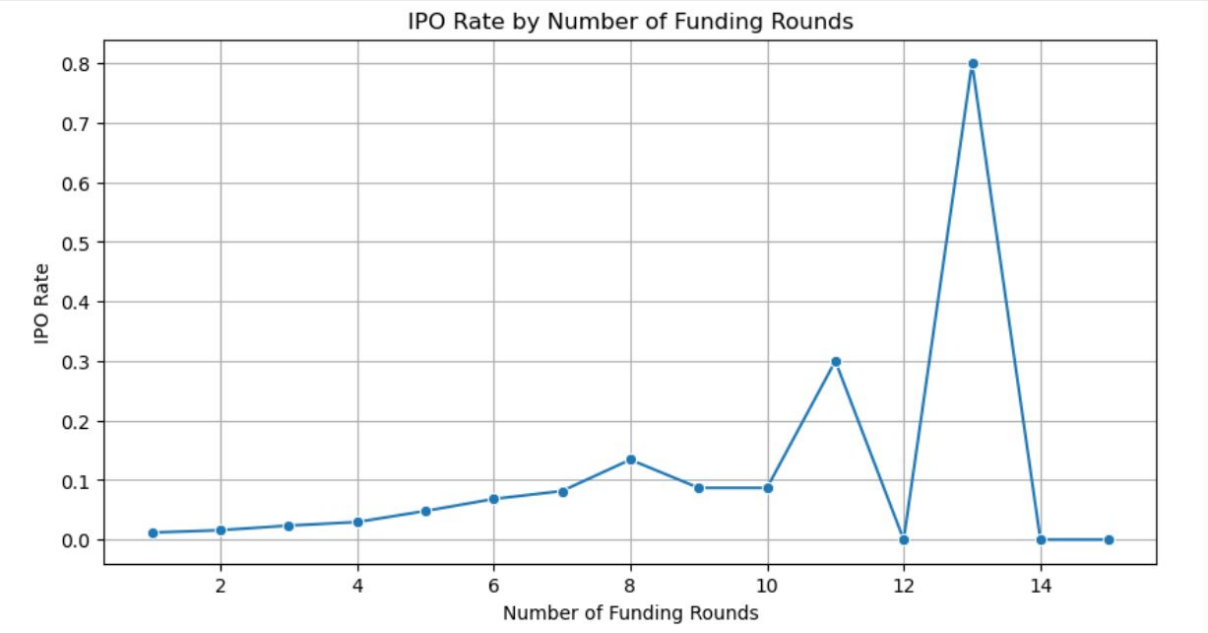
Data Analysis



Summary

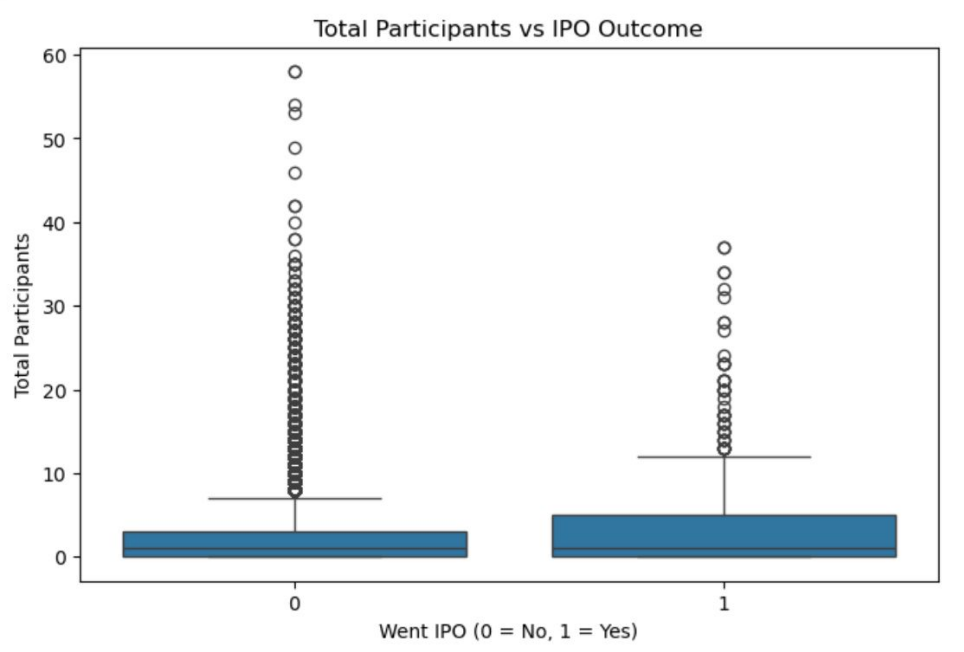
- Creates a histogram of total funding raised (log-transformed).
- Applies log(raised amount + 1) to handle skew and zero values.
- Uses 50 bins and overlays a KDE (density) curve.

Data Analysis



Data Analysis

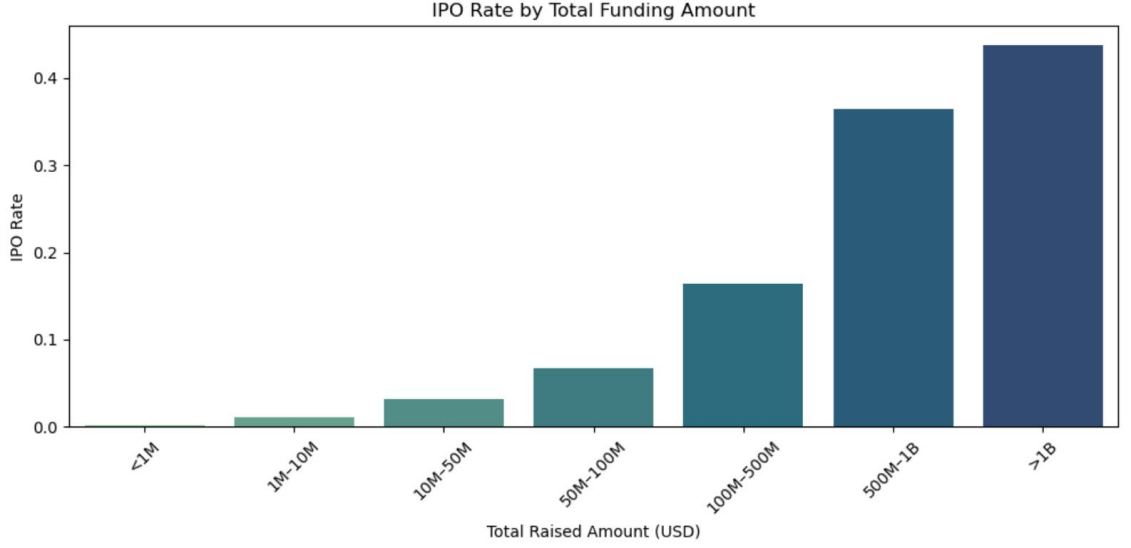
```
# Boxplot of Participants vs IPO Status
plt.figure(figsize=(8,5))
sns.boxplot(x='went_ipo_max', y='participants_sum', data=company_agg)
plt.title('Total Participants vs IPO Outcome')
plt.xlabel('Went IPO (0 = No, 1 = Yes)')
plt.ylabel('Total Participants')
plt.show()
```



Summary

- Creates a boxplot comparing participant counts by IPO status.
- Groups companies based on whether they went public or not.
- Shows distribution and outliers of total participants per group.

Data Analysis



Summary

- Categorizes companies into bins based on total funding raised.
- Assigns each company to its appropriate funding range.
- Computes the average IPO rate for each funding bin.

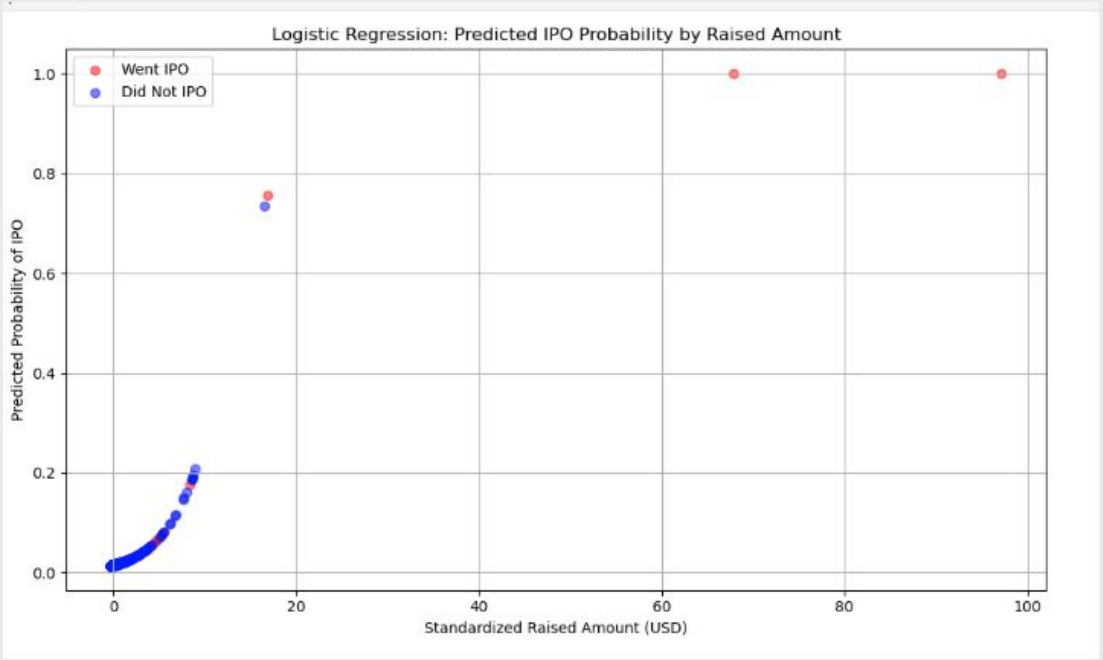
Data Analysis

```
Accuracy: 0.9838760175328741
Precision: 0.75
Recall: 0.02857142857142857
F1 Score: 0.05504587155963303

Classification Report:
              precision    recall  f1-score   support

           0       0.98      1.00      0.99      6283
           1       0.75      0.03      0.06       105

    accuracy                           0.98      6388
   macro avg       0.87      0.51      0.52      6388
weighted avg       0.98      0.98      0.98      6388
```

Summary

- Defines input features and IPO outcome.
- Scales features for better model performance.
- Splits data into training and testing sets.
- Trains a logistic regression model to predict IPO likelihood.
- Generates predictions and prediction probabilities.

Data Analysis

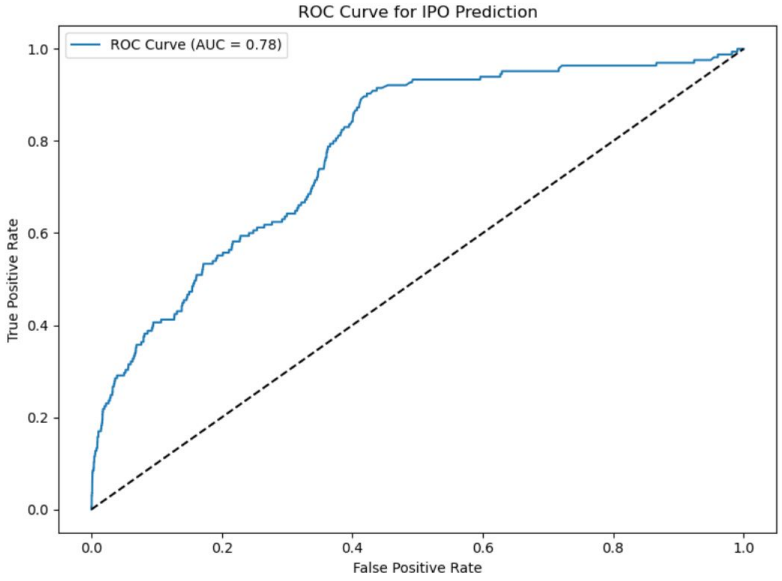


Summary

- Calculates and displays the ROC AUC score.
- Plots the ROC curve to evaluate model performance.

Data Analysis

```
Classification Report:
              precision    recall  f1-score   support

           0       0.98      1.00      0.99      9417
           1       0.38      0.03      0.06       165

    accuracy                           0.98      9582
   macro avg       0.68      0.51      0.52      9582
weighted avg       0.97      0.98      0.98      9582

Confusion Matrix:
[[9409    8]
 [ 160    5]]
ROC AUC Score: 0.780
```



Summary

- Calculates and displays the ROC AUC score.
- Plots the ROC curve to evaluate model performance.

Thanks for participating

Conclusion

1

Challenges

Data Merging Complexity: Matching records across datasets required careful alignment on unique identifiers which can be inconsistent or missing in real-world data.Complex

2

Challenges

Skewed Distributions Variables like total funding raised were highly skewed, requiring log transformation to support meaningful visualizations and modeling.

3

Suggestions

Add market conditions at time of funding IPO

Timing of funding rounds (years between) to capture growth

4

Suggestions

Interactive dashboards using Tableau to explore IPO trends and company features

Crunchbase API and Pitchbook

Thank you

Questions?