World's Billionaires' Statistics

MSIS 2607

Data Analytics Project

Contents

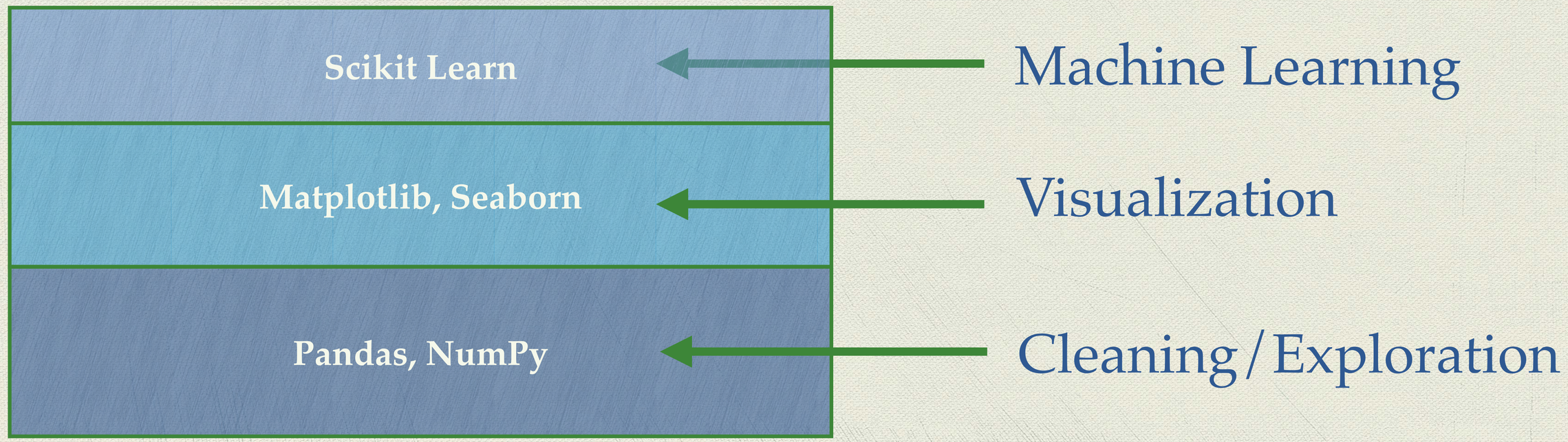
Overview

- The dataset describes the dynamics of wealth distribution, industries, and the personal details of the world's billionaires

- This project focusses on exploring the rankings, industries, economic indicators and demographics of billionaires and other influencing factors

- Further, conclusions are obtained based on the analysis

Libraries Used



Problem Statement

- Analyse and investigate the billionaires distribution across regions and industries and categorise them based on gender, spread of wealth to explore the geographical concentration of billionaires and it’s impact on wealth inequality

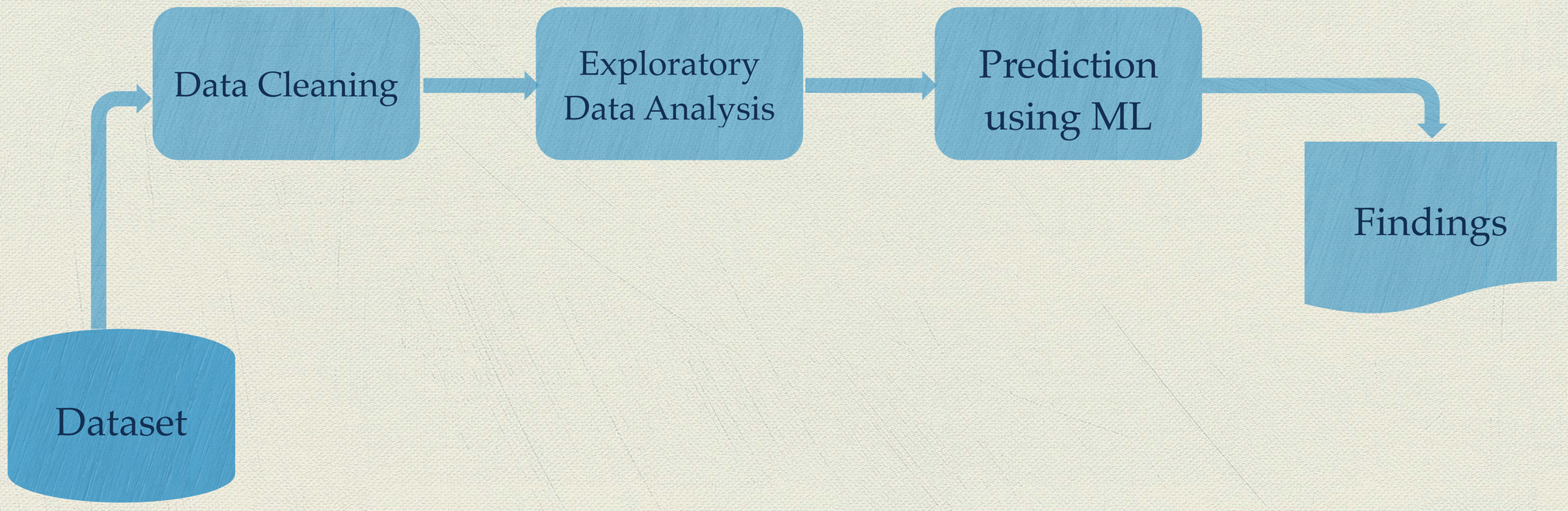
- Identify correlation(s) that may have impact on the rankings using strategic implementations and analysis to provide meaningful insights into the dynamics of the billionaires’ wealth

Dataset

- Shape – 2640 R x 35 C
- Each row is an entry of the information of one billionaire in the ranking table
- Details include : rank, name, age, net worth, industries, country, GDP etc.

rank	finalWorth	personName	age	country	city	source	industries	countryOfCitizenship	organization	selfMade	gender	lastName	firstName
1	211000	Bernard Arnault & family	74.0	France	Paris	LVMH	Fashion & Retail	France	LVMH Moët Hennessy Louis Vuitton	False	M	Arnault	Bernard
2	180000	Elon Musk	51.0	United States	Austin	Tesla, SpaceX	Automotive	United States	Tesla	True	M	Musk	Elon
3	114000	Jeff Bezos	59.0	United States	Medina	Amazon	Technology	United States	Amazon	True	M	Bezos	Jeff
4	107000	Larry Ellison	78.0	United States	Lanai	Oracle	Technology	United States	Oracle	True	M	Ellison	Larry
5	106000	Warren Buffett	92.0	United States	Omaha	Berkshire Hathaway	Finance & Investments	United States	Berkshire Hathaway Inc. (Cl A)	True	M	Buffett	Warren

Flow of Analysis



Data Cleaning

- Redundant columns were removed

```
df['category'].equals(df['industries'])
```

```
True
```

- Missing data was identified and addressed

```
df['country'].fillna('Unknown', inplace=True)
df['title'].fillna('Unknown', inplace=True)
df['organization'].fillna('Unknown', inplace=True)
df['residenceStateRegion'].fillna('Unknown', inplace=True)
df['state'].fillna('Unknown', inplace=True)
df['gdp_country'] = df['gdp_country'].str.replace('$', '').str.replace(',', '').astype(float)
```

Data Cleaning

- Columns irrelevant to the core of the analysis were dropped

1. birthDate
2. birthYear
3. birthMonth
4. birthDay

```
df.drop(columns = ['birthDate', 'birthYear', 'birthMonth', 'birthDay'], inplace = True)
```

```
df['date'].unique()
```

```
array(['4/4/2023 5:01', '4/4/2023 9:01'], dtype=object)
```

As the data in this dataset has been collected on the same day and since this information does not contribute to the overall analysis, the 'date' column is being dropped.

```
df.drop(columns = 'date', inplace = True)
```

```
# Display the unique values present in the 'status' column
df['status'].unique()
```

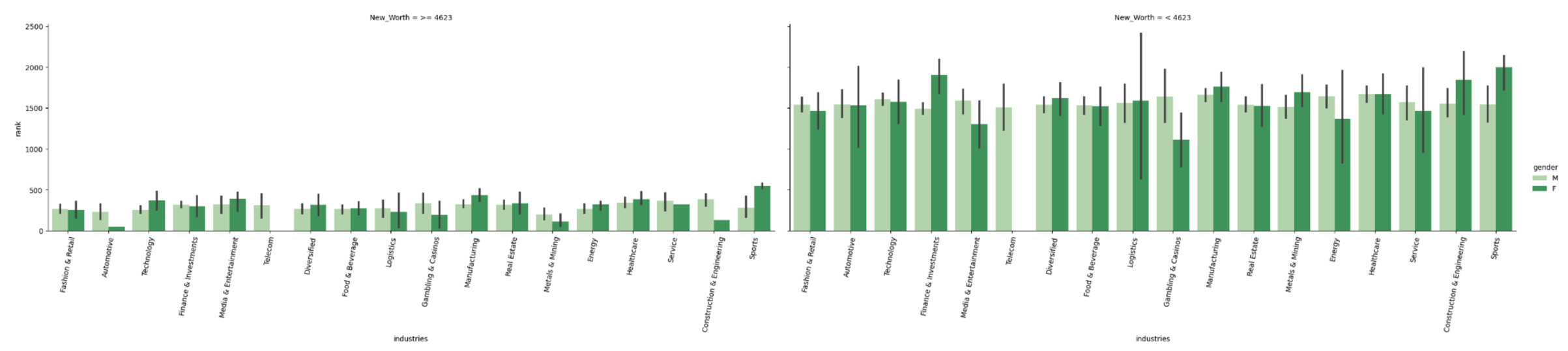
```
array(['U', 'D', 'N', 'Split Family Fortune', 'E', 'R'], dtype=object)
```

```
# Drop the 'status' column from the DataFrame
df.drop(columns = 'status', inplace = True)
```

Finding 1 + Insights

Rank, Industries and Gender Analysis w.r.t Net Worth

- The net worth was categorised based on mean
- For each net worth category, the rank vs industry comparison was analysed
- Above the mean net worth, the male billionaires outperform the female billionaires. However, below the mean net worth, the ranking variance is lower



Finding 1

Rank, Industries and Gender Analysis w.r.t Net Worth

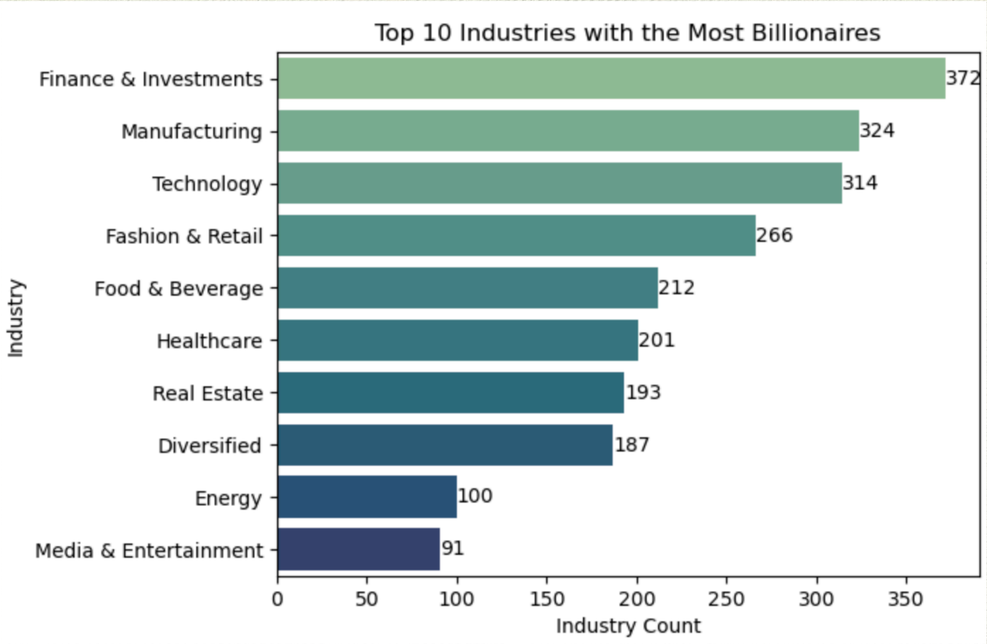
```
# Calculating the mean of the 'finalWorth' column
final_worth_mean = df.finalWorth.mean()

# Creating a new column 'New_Worth' based on a condition
df['New_Worth'] = df.finalWorth >= final_worth_mean

# Replacing True values with '>= final_worth_mean' and False values with '< final_worth_mean'
df.New_Worth.replace(to_replace=True, value='>= %d' % final_worth_mean, inplace=True)
df.New_Worth.replace(to_replace=False, value='< %d' % final_worth_mean, inplace=True)

# Creating a categorical plot using Seaborn
g = sns.catplot(y='rank', x='industries', hue='gender', col='New_Worth', data=df, kind='bar', aspect=3, palette="Gre

# Rotating x-axis labels for better visibility
g.set_xticklabels(rotation=80)
```

Finding 2 (a) + Insights

Billionaires Industry-wise + Wealth Distribution by Industry

- The following plot shows that Finance & Investments category produces the maximum number of billionaires.

- Assumption: Finance & Investments will now have the maximum wealth distribution by industry, followed by Manufacturing and Technology.

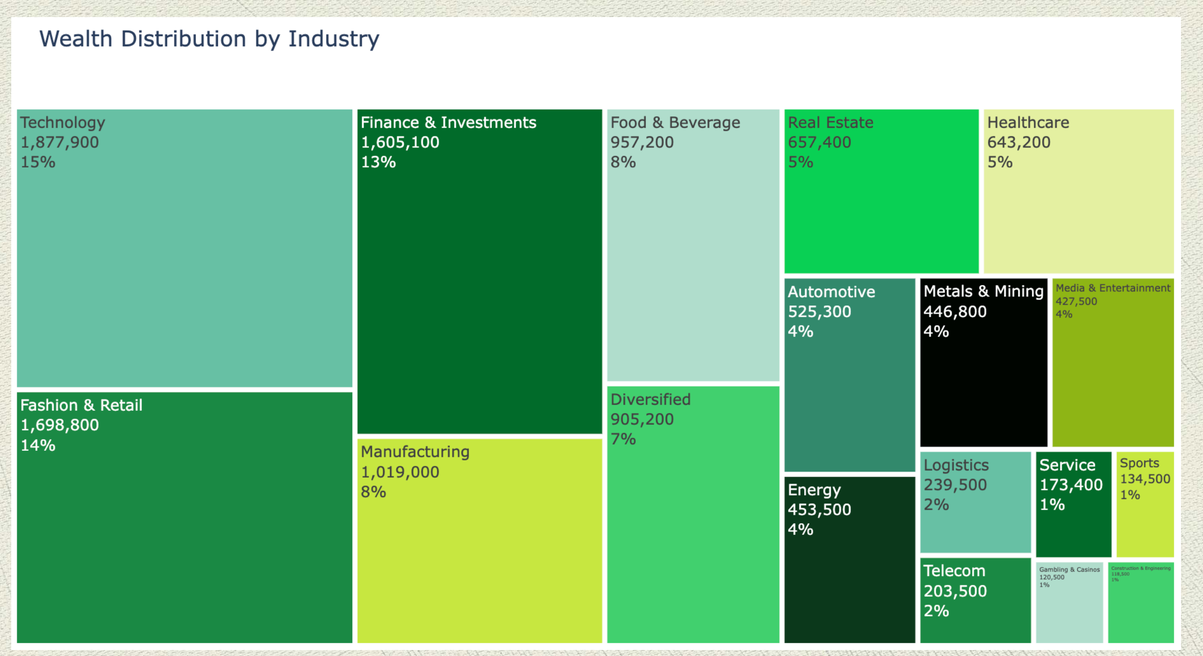


Finding 2 (b) + Insights

Billionaires Industry-wise + Wealth Distribution by Industry

- The treemap tells us that Technology owns the most(15%) stake in Wealth Distribution. Though Finance & Investments produced the maximum number of billionaires, the net wealth they brought in was still behind that of Technology.

- Conclusion: Maximum number of billionaires in an industry does not imply maximum wealth distribution of that industry vertical



Finding 2 (a & b)

Billionaires Industry-wise + Wealth Distribution by Industry

```python
# Extracting the top 10 industries with the most billionaires and their counts
e = df['industries'].value_counts().head(10)

# Creating a bar plot using Seaborn
c_plot = sns.barplot(y=e.index, x=e.values, palette = "crest")

# Adding labels and title to the plot
plt.xlabel('Industry Count')  # X-axis label
plt.ylabel('Industry')  # Y-axis label
plt.title('Top 10 Industries with the Most Billionaires')  # Plot title

# Rotating x-axis labels for better readability
c_plot.set_xticklabels(c_plot.get_xticklabels(), rotation=360)

# Adding labels to the bars in the plot
for i in c_plot.containers:
    c_plot.bar_label(i,)

# Displaying the plot
plt.show()
```

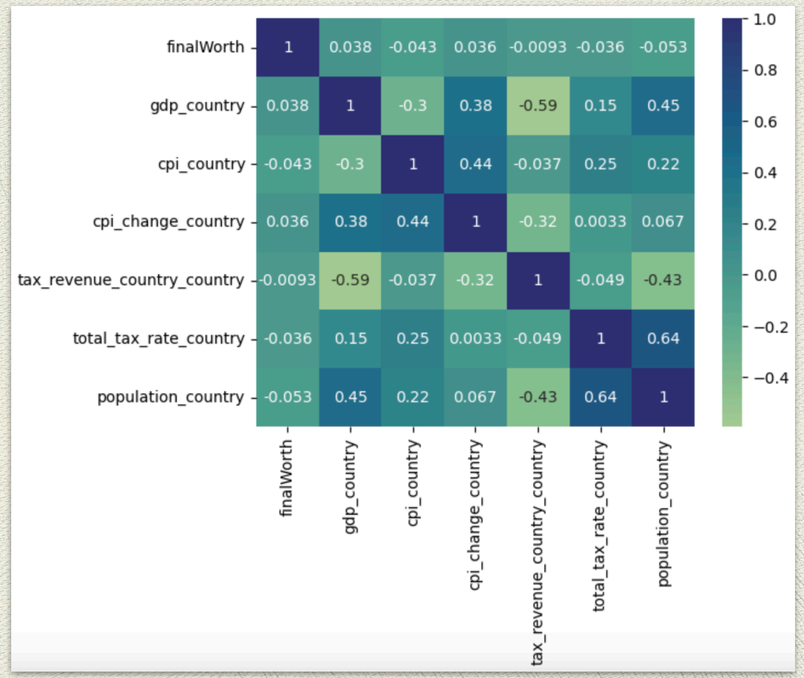
```python
# Defining the custom color palette
custom_green_palette = [
    '#66c2a5',  # A lighter green
    '#238b45',  # A medium green
    '#006d2c',  # A darker green
    '#C7EA46'   # An even darker green
]

# Creating a treemap using Plotly Express
fig7 = px.treemap(df,
                  path=['industries'], values='finalWorth',
                  title='Wealth Distribution by Industry',
                  color_discrete_sequence=custom_green_palette)

# Updating traces to display additional information on the treemap
fig7.update_traces(textinfo="label+percent entry+value")
fig7.update_traces(hovertemplate='labels = %{label}<br>finalWorth = $%{value}<extra></extra>')

# Updating layout to customize the margin
fig7.update_layout(margin=dict(t=50, l=25, r=25, b=25))

# Displaying the treemap
fig7.show()
```

Finding 3 + Insights

Correlation between the Wealth of Billionaires and National Economic Indicators

- **Total Tax Rate and Population Correlation: 0.644**
  Indicates the strong correlation, that larger countries tend to have higher total tax rates.. , no impact on billionaires status

- **Final Worth and Population (population_country) Correlation: -0.053**
  A weak inverse correlation suggests that billionaires tend to have slightly less wealth in countries with larger populations.

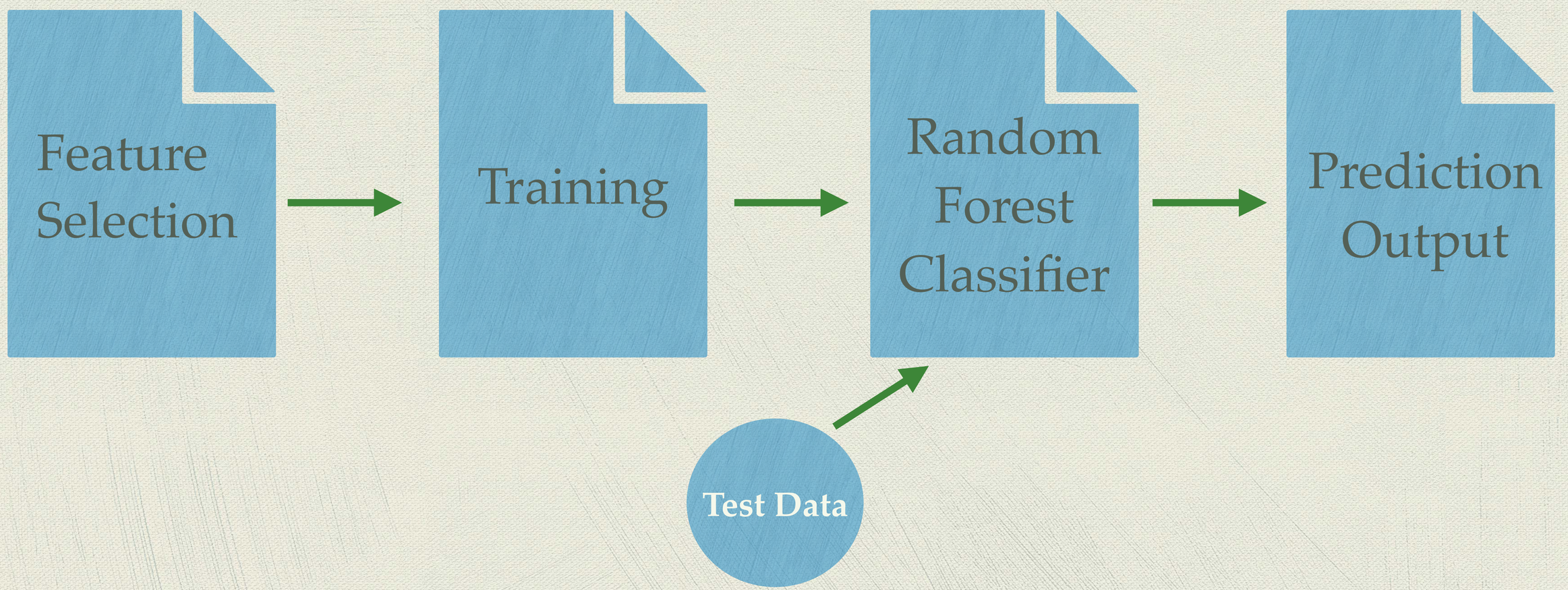


Finding 3

Correlation between the Wealth of Billionaires and National Economic Indicators

```
# Selecting columns for correlation analysis
correlation_data = df[[
    'finalWorth',
    'gdp_country',
    'cpi_country',
    'cpi_change_country',
    'tax_revenue_country_country',
    'total_tax_rate_country',
    'population_country'
]]
# Creating a heatmap to visualize the correlation matrix
sns.heatmap(correlation_data.corr(),annot=True,cmap="crest")
ax.set(xlabel="", ylabel="")
```

Machine Learning



Machine Learning

Feature Selection

```
# Feature Selection and Data Cleaning
# Exclude specific columns from the original DataFrame 'df' to create the feature DataFrame 'X'
X = df.drop(['industries', 'title','state','residenceStateRegion',"city",'lastName', 'firstName',
             'latitude_country','longitude_country', 'personName', 'gdp_country','New_Worth','organization'], axis =

# Replace missing values (NaN) in the feature DataFrame 'X' with the value 0.0
X.replace(to_replace=np.nan, value =0.0, inplace=True)

# Target Variable
# Create a Series 'y' representing the target variable ('industries') for machine learning
y =df['industries']

# initializes an empty dictionary named label_encoders to store individual LabelEncoder objects for each feature.
label_encoders = {}
features = ['rank', 'gender','country', 'countryOfCitizenship', 'selfMade', 'source']
for feature in features:
    label_encoders[feature] = LabelEncoder()
    X[feature] = label_encoders[feature].fit_transform(X[feature])
X.head()
```

Machine Learning

Training / Testing

```
# Train-Test Split
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.3,random_state = 0)
```

```
rfc = RandomForestClassifier(random_state = 0)

rfc.fit(X_train,y_train)
```

```
▾        RandomForestClassifier
RandomForestClassifier(random_state=0)
```

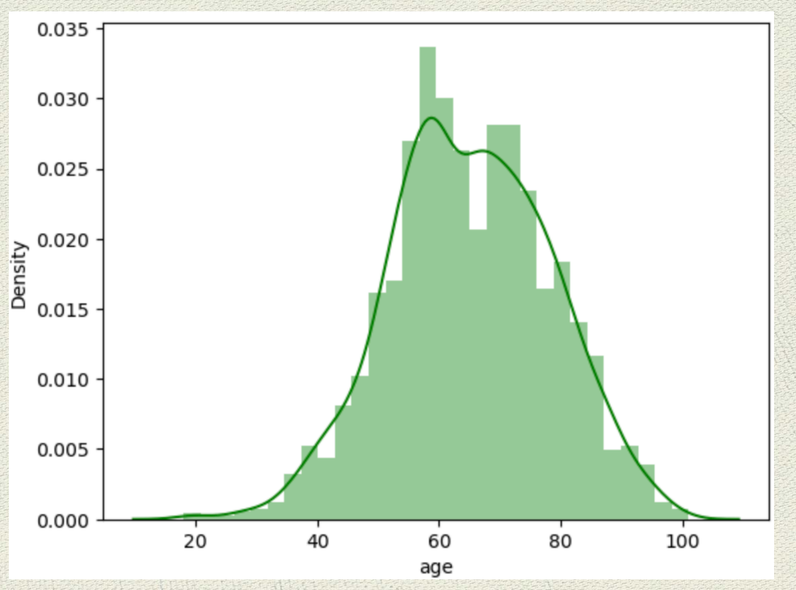
```
# 'X_test' contains the features for which predictions are made, and 'y_pred' stores the predicted values.
y_pred = rfc.predict(X_test)
```

```
In [54]: y_pred
```

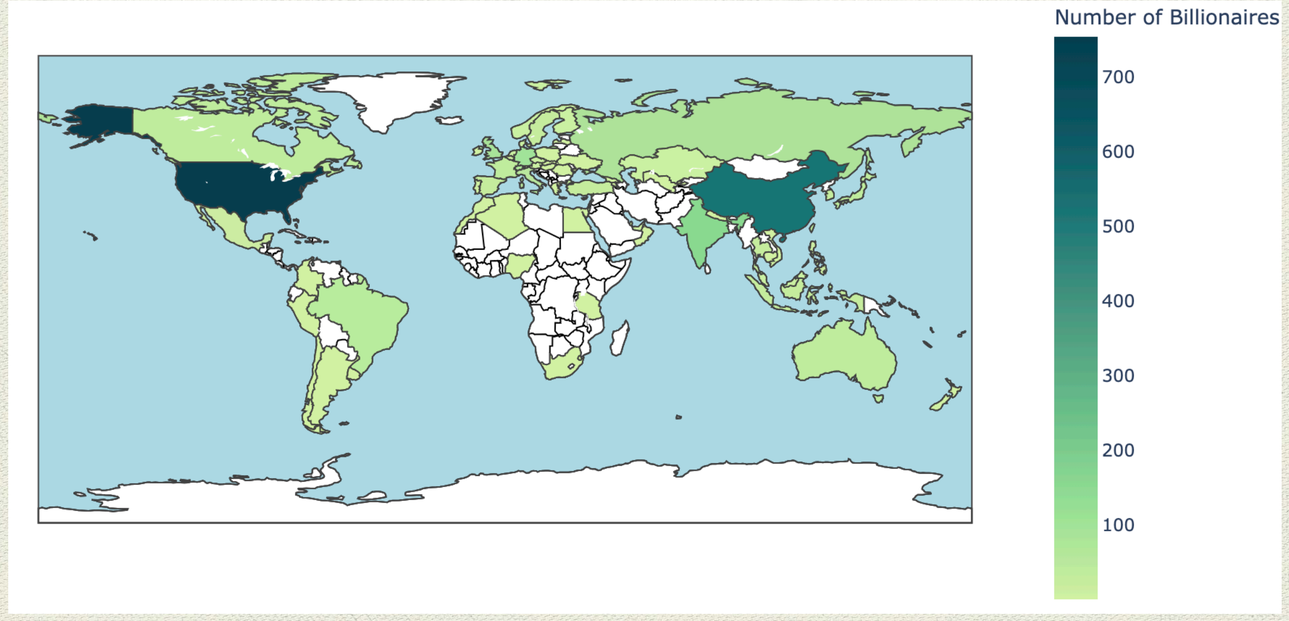
```
Out[54]: array(['Manufacturing', 'Finance & Investments', 'Finance & Investments',
                'Technology', 'Technology', 'Manufacturing', 'Healthcare',
                'Energy', 'Finance & Investments', 'Manufacturing',
                'Finance & Investments', 'Manufacturing', 'Technology',
                'Technology', 'Technology', 'Fashion & Retail', 'Diversified',
                'Manufacturing', 'Energy', 'Real Estate', 'Energy', 'Technology',
                'Manufacturing', 'Manufacturing', 'Diversified', 'Manufacturing',
                'Technology', 'Diversified', 'Real Estate', 'Sports',
```

Additional Findings 1

Billionaires age distribution

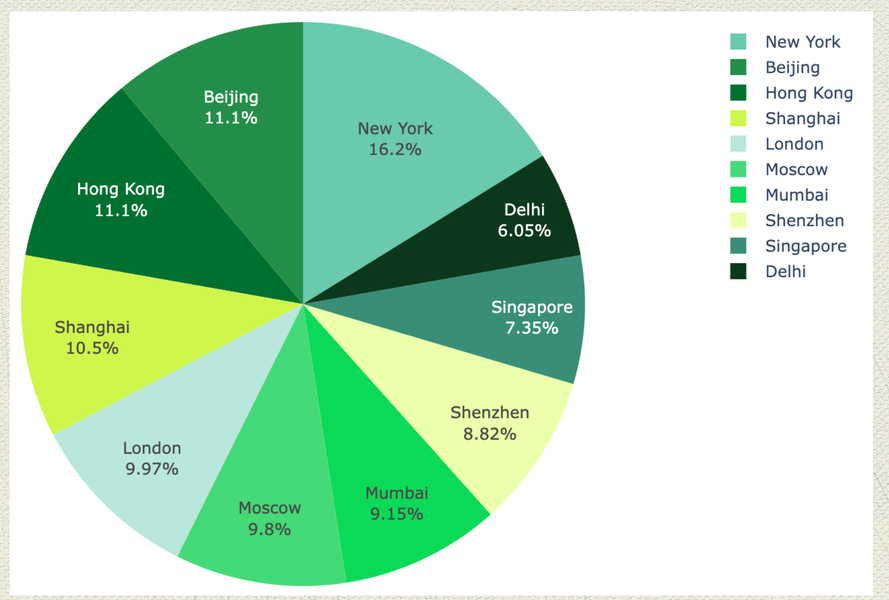


Billionaires concentration across the world

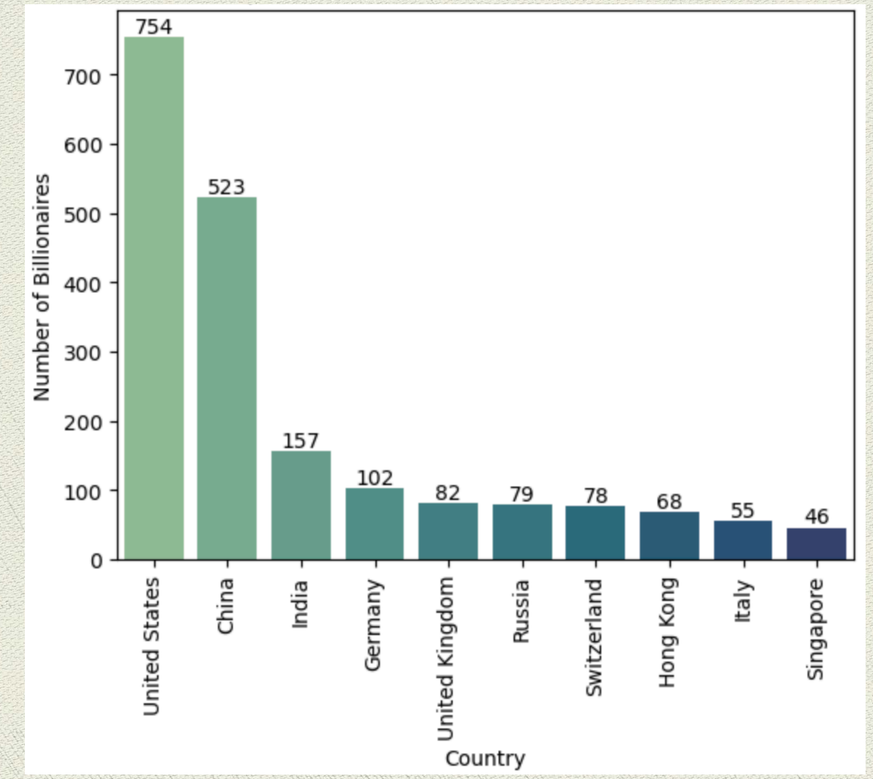


Additional Findings 2

Cities with the most billionaires

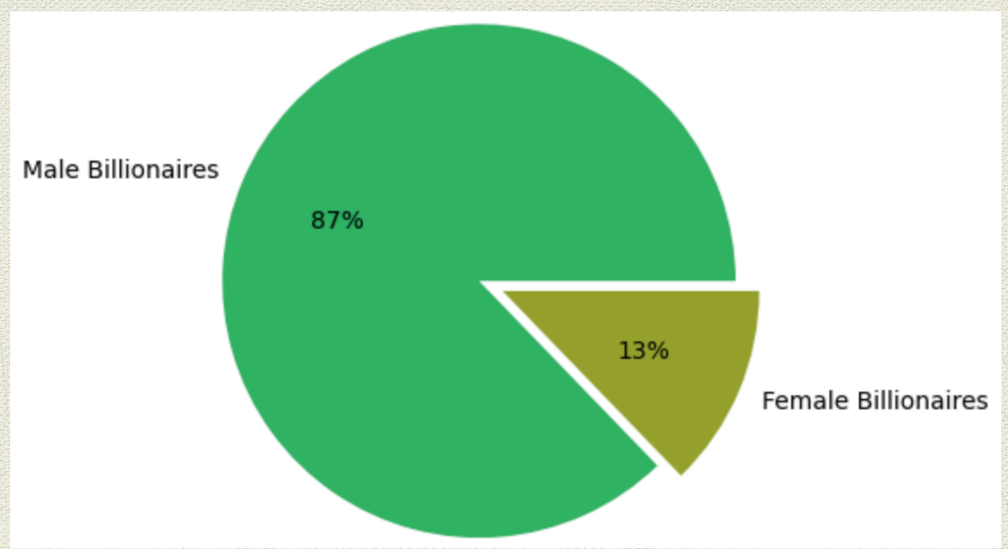


Countries with most billionaires

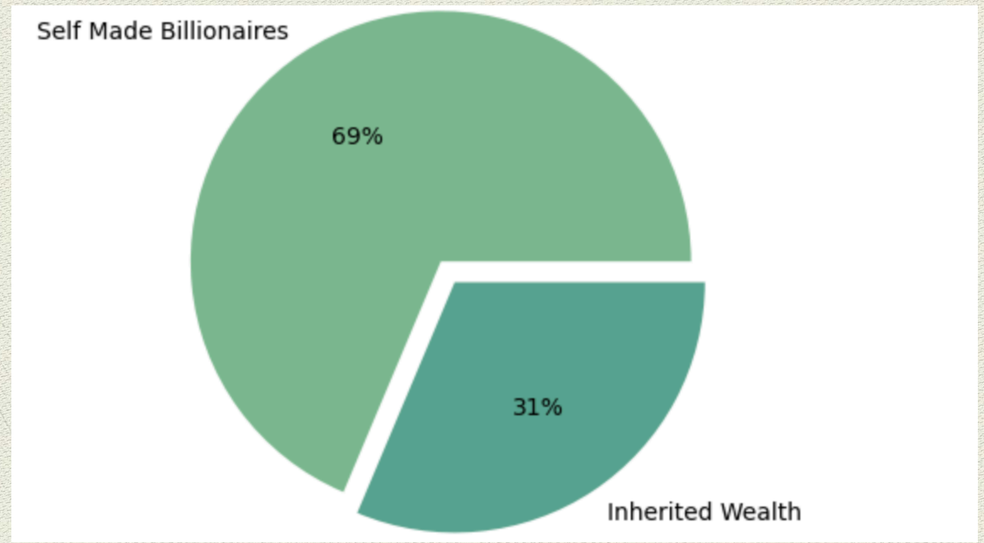


Additional Findings 3

Distribution of billionaires, w.r.t gender

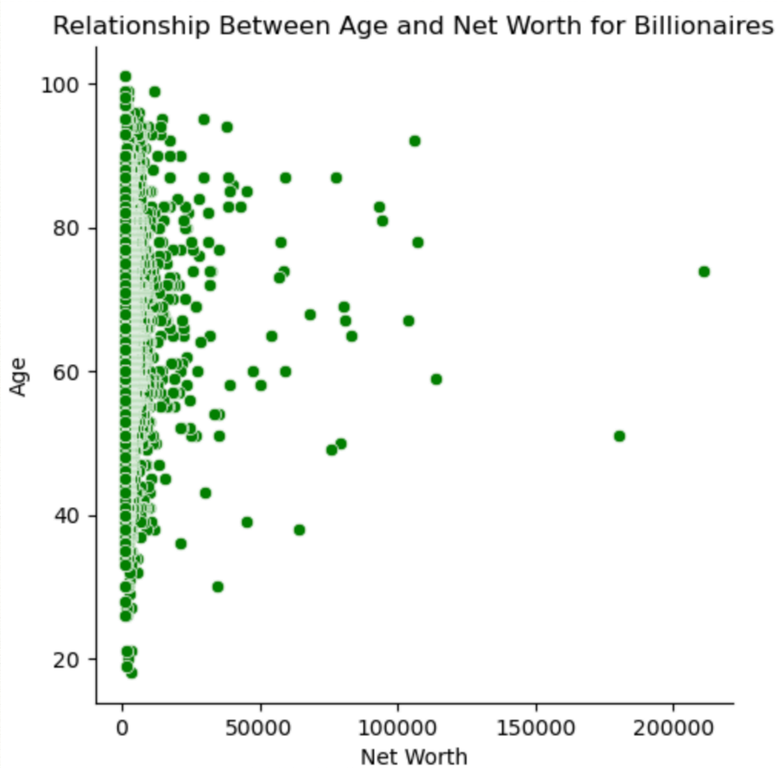


Billionaires classification based on inheritance

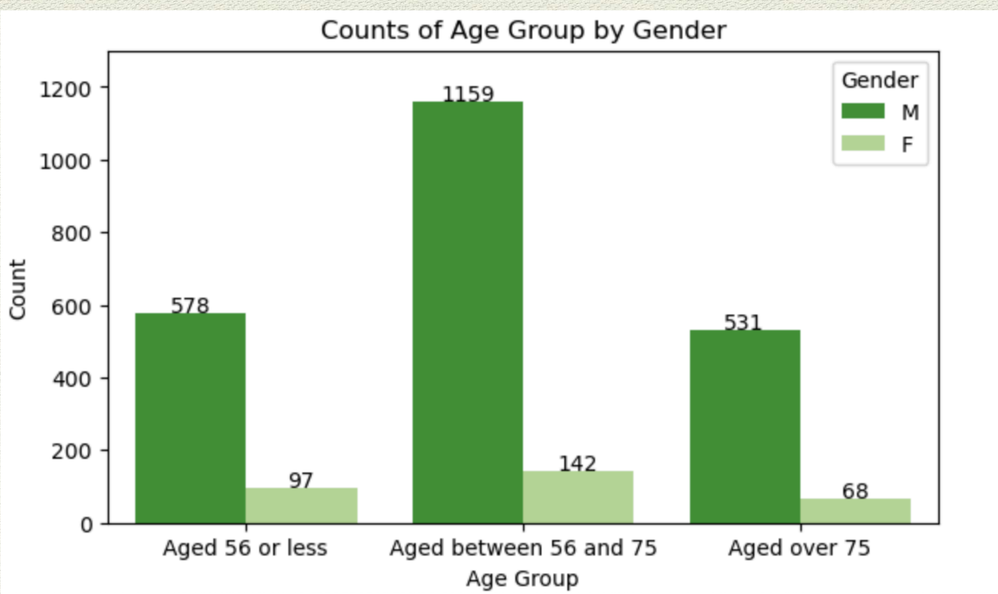


Additional Findings 4

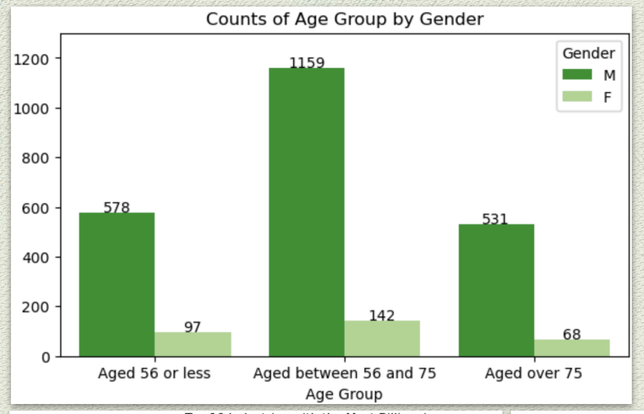
Age vs Net Worth of billionaires

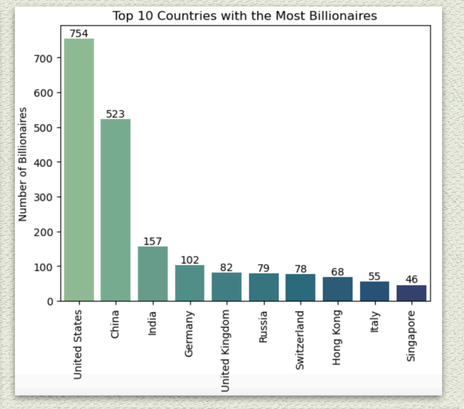


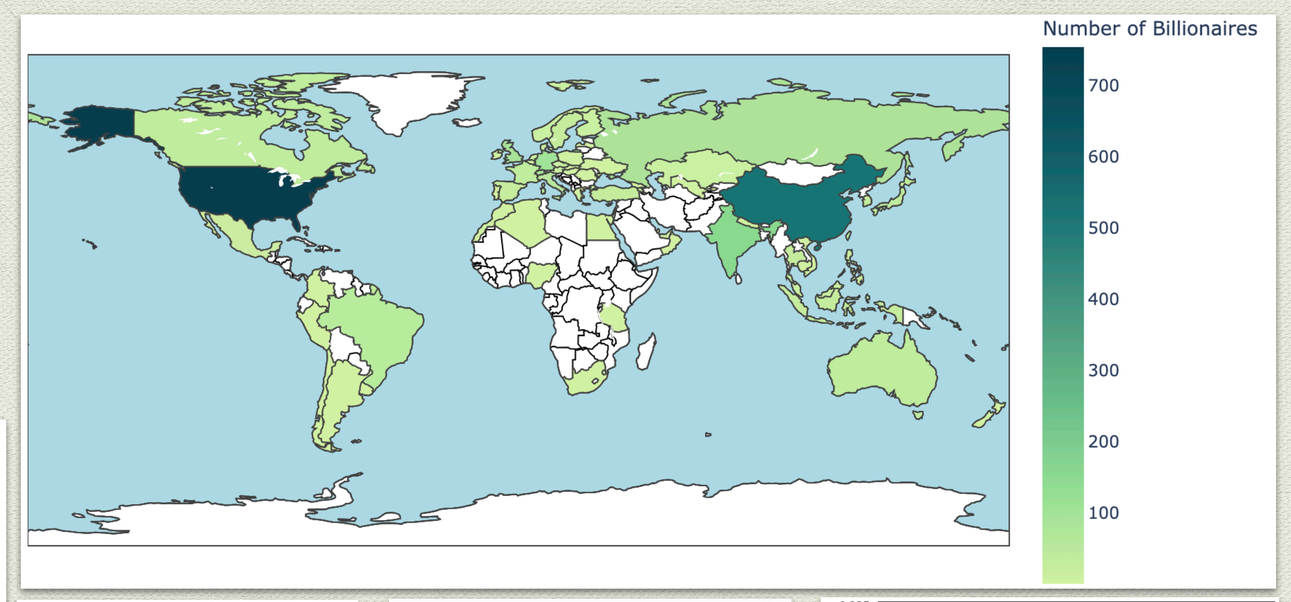
Relationship between rank and net worth, segmented by gender

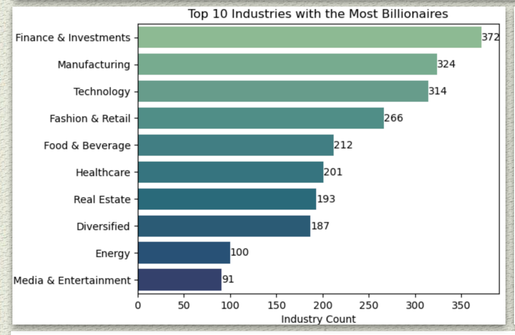


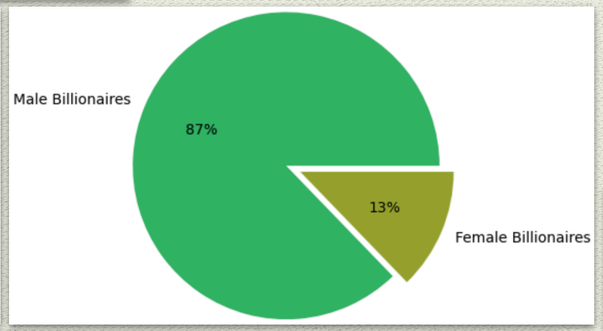
Thank You!

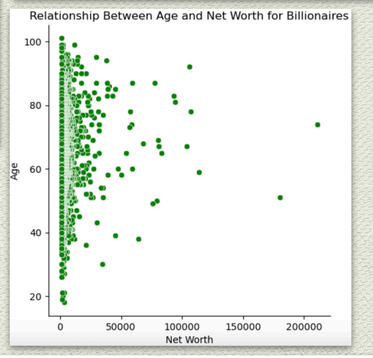

Counts of Age Group by Gender
Gender
M
F
578
97
1159
142
531
68
Aged 56 or less
Aged between 56 and 75
Aged over 75
Age Group

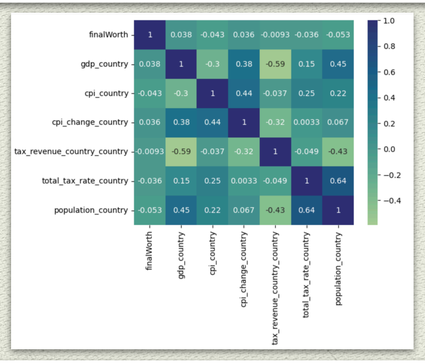

Top 10 Countries with the Most Billionaires
Number of Billionaires
754
523
157
102
82
79
78
68
55
46
United States
China
India
Germany
United Kingdom
Russia
Switzerland
Hong Kong
Italy
Singapore

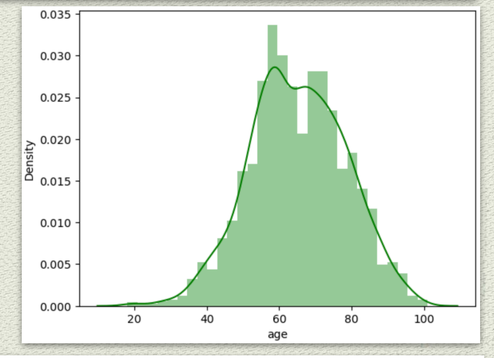

Number of Billionaires
700
600
500
400
300
200
100

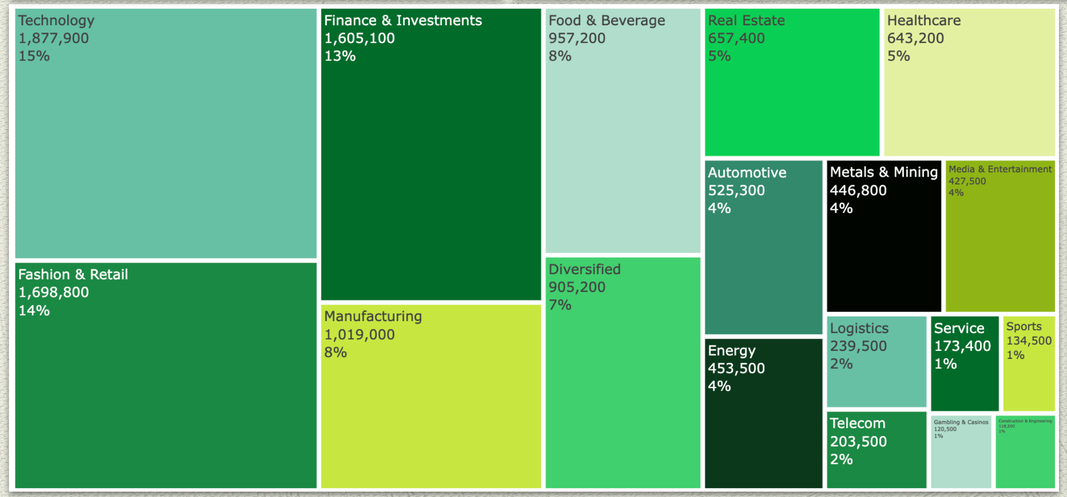

Top 10 Industries with the Most Billionaires
Finance & Investments
Manufacturing
Technology
Fashion & Retail
Food & Beverage
Healthcare
Real Estate
Diversified
Energy
Media & Entertainment
372
324
314
266
212
201
193
187
100
91
Industry Count

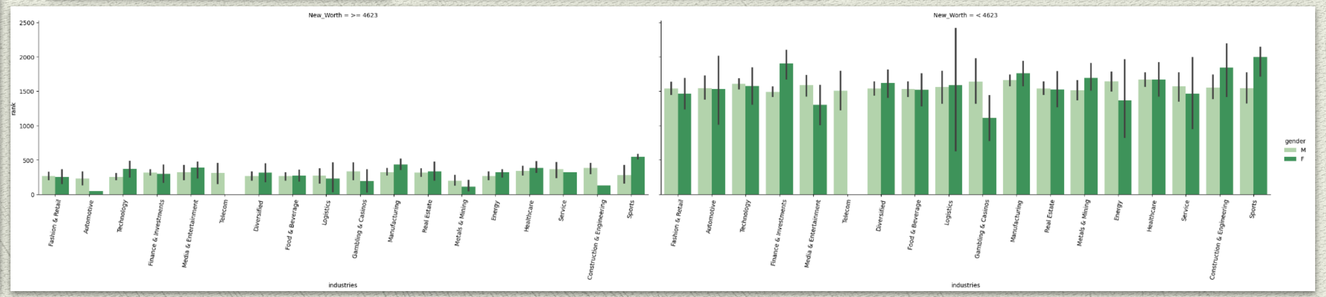

Male Billionaires
87%
13%
Female Billionaires


Relationship Between Age and Net Worth for Billionaires
Age
Net Worth


finalWorth
gdp_country
cpi_country
cpi_change_country
tax_revenue_country_country
total_tax_rate_country
population_country


Density
age


Technology
1,877,900
15%
Finance & Investments
1,605,100
13%
Food & Beverage
957,200
8%
Real Estate
657,400
5%
Healthcare
643,200
5%
Fashion & Retail
1,698,800
14%
Manufacturing
1,019,000
8%
Diversified
905,200
7%
Automotive
525,300
4%
Metals & Mining
446,800
4%
Media & Entertainment
427,500
4%
Energy
453,500
4%
Logistics
239,500
2%
Service
173,400
1%
Sports
134,500
1%
Telecom
203,500
2%


New_Worth = >= 4623
New_Worth = < 4623
industries
gender
M
F