UK_Industry_ProgrammingLang_Analysis

March 17, 2025

[ ]:
```
import pandas as pd
from IPython.display import display
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.cluster import KMeans
from sklearn.metrics import classification_report
```

[20]:
```
df = pd.read_csv('survey_results_public.csv')

# Filter DataFrame to only include entries from the UK
df_uk = df[df['Country'] == 'United Kingdom of Great Britain and Northern␣
 ↪Ireland']
missing_industry = df_uk['Industry'].isnull().sum() / len(df_uk) * 100
missing_languages = df_uk['LanguageHaveWorkedWith'].isnull().sum() / len(df_uk)␣
 ↪* 100

print(f"Industry missing rate: {missing_industry:.2f}%")
print(f"LanguageHaveWorkedWith missing rate: {missing_languages:.2f}%")

# print(df_uk.head())

print(f"Total UK entries: {df_uk.shape[0]}")
```

```
Industry missing rate: 50.81%
LanguageHaveWorkedWith missing rate: 1.55%
Total UK entries: 3224
```

0.0.1 Displaying a Sample of the UK Dataset

Removes NaN values to ensure that the dataset sample is clean for presentation.

[23]:
```
df_uk_cleaned = df_uk.dropna(subset=['Industry', 'LanguageHaveWorkedWith'])

selected_columns = df_uk_cleaned[['Industry', 'LanguageHaveWorkedWith']].head()

display(selected_columns)
```

```
                      Industry  \
45                      Energy
54         Software Development
86         Software Development
100                     Fintech
133  Banking/Financial Services

                                LanguageHaveWorkedWith
45             C#;Dart;Fortran;Go;Julia;Python;SQL;VBA
54   C++;HTML/CSS;JavaScript;Kotlin;Lua;PowerShell;…
86   Bash/Shell (all shells);C#;HTML/CSS;JavaScript…
100                    Bash/Shell (all shells);Go;SQL
133  Bash/Shell (all shells);C#;HTML/CSS;Java;JavaS…
```

0.0.2 Handling Missing Industry Data

Since the 'Industry' field has 50% missing values, replacing NaN with 'Unknown' allows us to retain more data for analysis.

[26]:
```
df_uk = df_uk.copy()
df_uk['Industry'] = df_uk['Industry'].fillna('Unknown')
df_uk_cleaned = df_uk.dropna(subset=['LanguageHaveWorkedWith'])
```

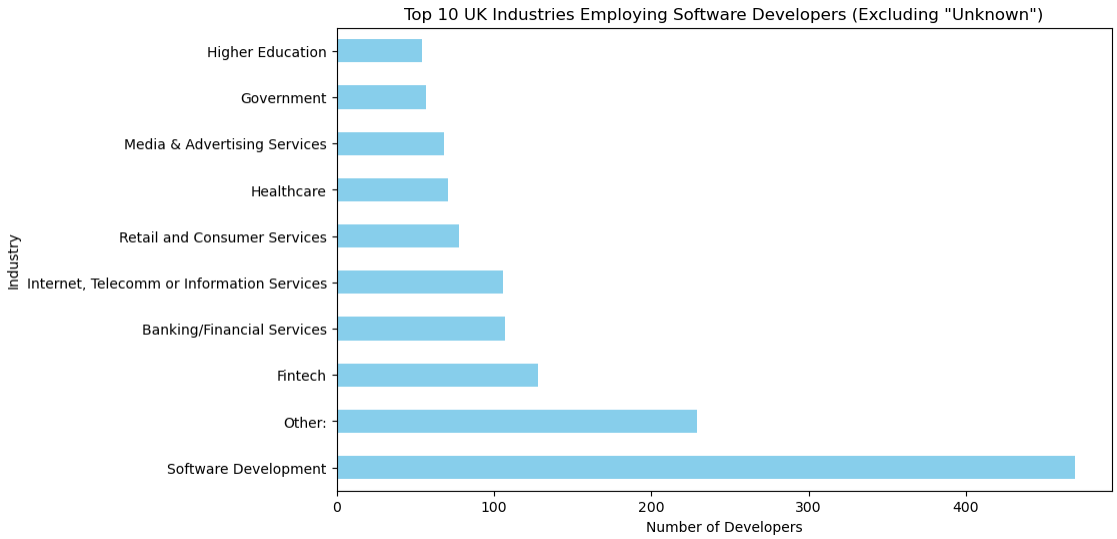
0.0.3 Analysing Industry Distribution

The analysis in this study is based on the Stack Overflow Annual Developer Survey 2024, which represents industries that currently employ significant numbers of software developers. Although it does not directly quantify open job vacancies or immediate hiring demand, it provides valuable insights into industries with long-term demand and established opportunities for software engineers in the UK.

[29]:
```
df_valid_industries = df_uk_cleaned[df_uk_cleaned['Industry'] != 'Unknown']

industry_counts = df_valid_industries['Industry'].value_counts()

print(industry_counts)

top_10_industries = industry_counts.head(10)

plt.figure(figsize=(10, 6))
top_10_industries.plot(kind='barh', color='skyblue')
plt.title('Top 10 UK Industries Employing Software Developers (Excluding␣
 ↪"Unknown")')
plt.xlabel('Number of Developers')
plt.ylabel('Industry')
plt.show()
```

```
Industry
Software Development                    469
```

```
Other:                                          229
Fintech                                         128
Banking/Financial Services                      107
Internet, Telecomm or Information Services      106
Retail and Consumer Services                     78
Healthcare                                       71
Media & Advertising Services                     68
Government                                       57
Higher Education                                 54
Manufacturing                                    52
Energy                                           50
Computer Systems Design and Services             47
Transportation, or Supply Chain                  46
Insurance                                        16
Name: count, dtype: int64
```

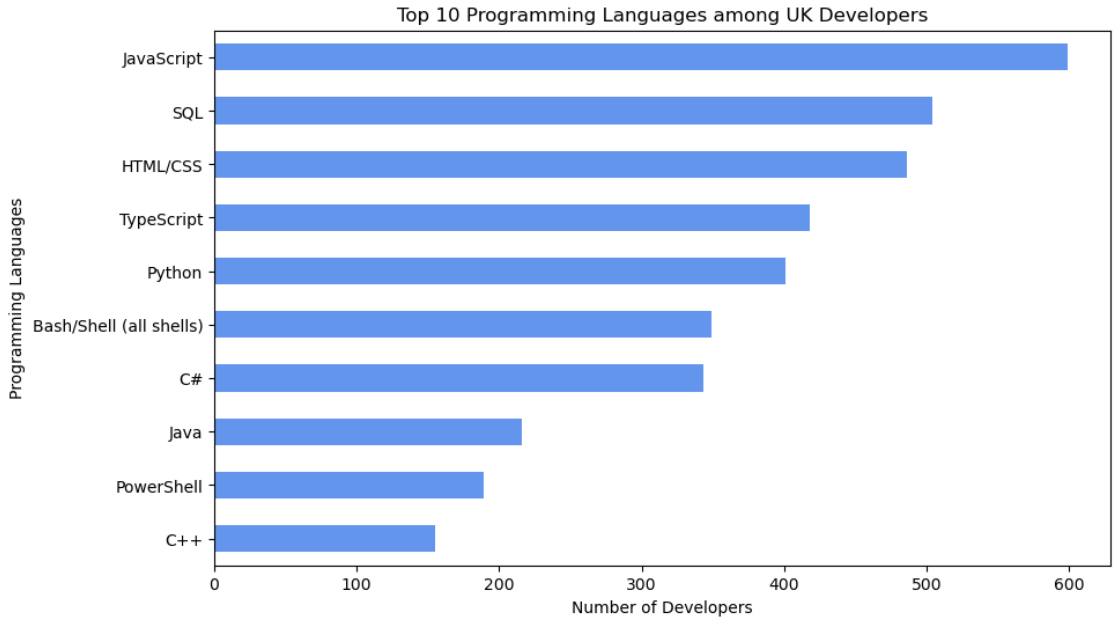


0.0.4 Top 10 Programming Languages in Top UK Industries

Determine which programming languages are most in demand across these sectors, helping prospective job seekers align their skills with market needs.

[32]:
```
df_uk_cleaned = df_uk_cleaned.copy()
df_uk_cleaned.loc[:, 'Industry'] = df_uk_cleaned['Industry'].str.strip()

top_industries = ['Software Development', 'Fintech', 'Banking/Financial␣
 ↪Services',
                  'Internet, Telecomm or Information Services', 'Retail and␣
 ↪Consumer Services']
```

```
df_top5 = df_uk_cleaned[df_uk_cleaned['Industry'].isin(top_industries)].copy()

df_top5 = df_top5.dropna(subset=['LanguageHaveWorkedWith'])

languages_series = df_top5['LanguageHaveWorkedWith'].str.split(';',␣
↪expand=True).stack()

language_counts = languages_series.value_counts().head(10)

plt.figure(figsize=(10,6))
language_counts.plot(kind='barh', color='cornflowerblue')
plt.title('Top 10 Programming Languages among UK Developers')
plt.xlabel('Number of Developers')
plt.ylabel('Programming Languages')
plt.gca().invert_yaxis()
plt.show()
```

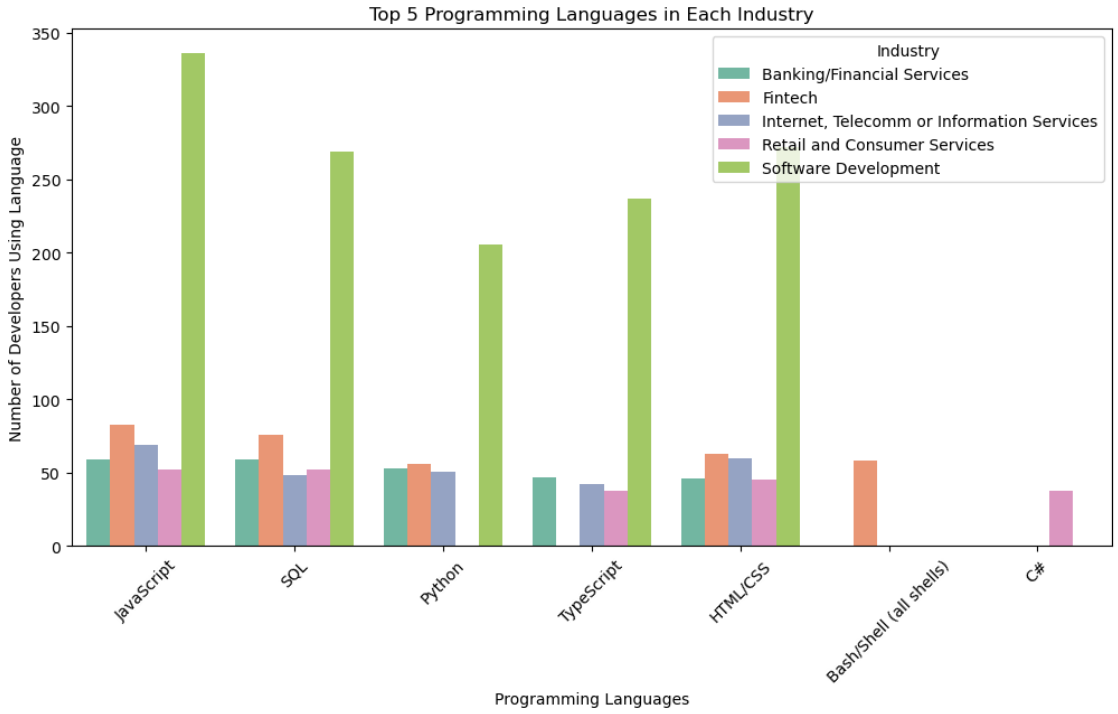


0.0.5 Analysing Industry-Specific Language Preferences

Determine which languages are most prevalent within each sector, allowing job seekers to understand sector-specific language preferences.

[35]:
```
df_languages = df_top5['LanguageHaveWorkedWith'].str.split(';', expand=True).
↪stack().reset_index(level=1, drop=True)
df_languages.name = 'Programming Language'
```

```
df_top5_lang = df_top5[['Industry']].join(df_languages)

top_languages_by_industry = df_top5_lang.groupby('Industry')['Programming␣
↪Language'].value_counts()

top_languages_df = top_languages_by_industry.groupby(level=0).head(5).
↪reset_index(name='Count')

plt.figure(figsize=(12, 6))

sns.barplot(data=top_languages_df, x="Programming Language", y="Count",␣
↪hue="Industry", palette="Set2")

plt.title('Top 5 Programming Languages in Each Industry')
plt.xlabel('Programming Languages')
plt.ylabel('Number of Developers Using Language')
plt.xticks(rotation=45)
plt.legend(title="Industry")
plt.show()
```



0.0.6 Random Forest

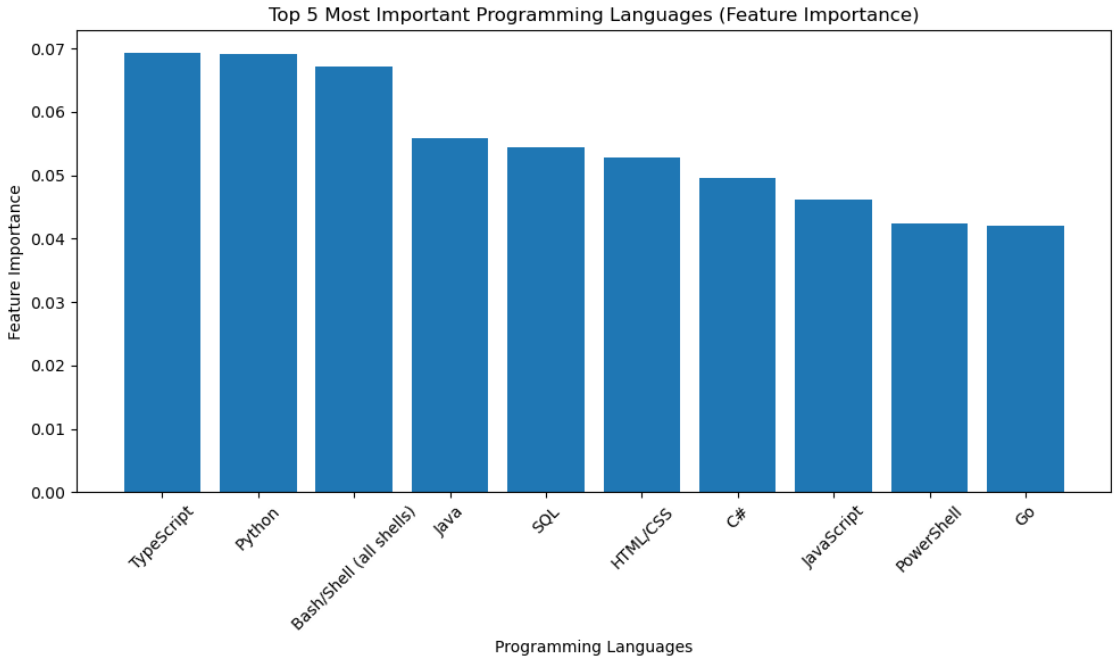
This analysis utilises data from the Stack Overflow 2024 survey, focusing on developers in the UK. Selected Top5 industries (Software Development, Fintech, Banking/Financial Services, Internet,

Telecomm or Information Services, and Retail and Consumer Services) and used programming languages (LanguageHaveWorkedWith) as features to explore associations with industry. The Random Forest method was employed to identify the most influential languages.

[42]:
```
languages = df_top5['LanguageHaveWorkedWith'].str.split(';')
df_languages = pd.get_dummies(languages.apply(pd.Series).stack()).
 ↪groupby(level=0).sum()

X = df_languages
y = df_top5['Industry']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,␣
 ↪random_state=42)

rf_model = RandomForestClassifier(n_estimators=100, class_weight='balanced',␣
 ↪random_state=42)
rf_model.fit(X_train, y_train)

language_names = df_languages.columns
importances = rf_model.feature_importances_
feature_importance = list(zip(language_names, importances))
top_10 = sorted(feature_importance, key=lambda x: x[1], reverse=True)[:10]

print("Top 10 most important languages:")
for lang, imp in top_10:
    print(f"{lang}: {imp:.4f}")

y_pred = rf_model.predict(X_test)
print(classification_report(y_test, y_pred))

top_10_langs, top_10_importances = zip(*top_10)
plt.figure(figsize=(10, 6))
plt.bar(top_10_langs, top_10_importances)
plt.title("Top 5 Most Important Programming Languages (Feature Importance)")
plt.xlabel("Programming Languages")
plt.ylabel("Feature Importance")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

```
Top 10 most important languages:
TypeScript: 0.0694
Python: 0.0690
Bash/Shell (all shells): 0.0671
Java: 0.0558
SQL: 0.0545
HTML/CSS: 0.0529
C#: 0.0495
```

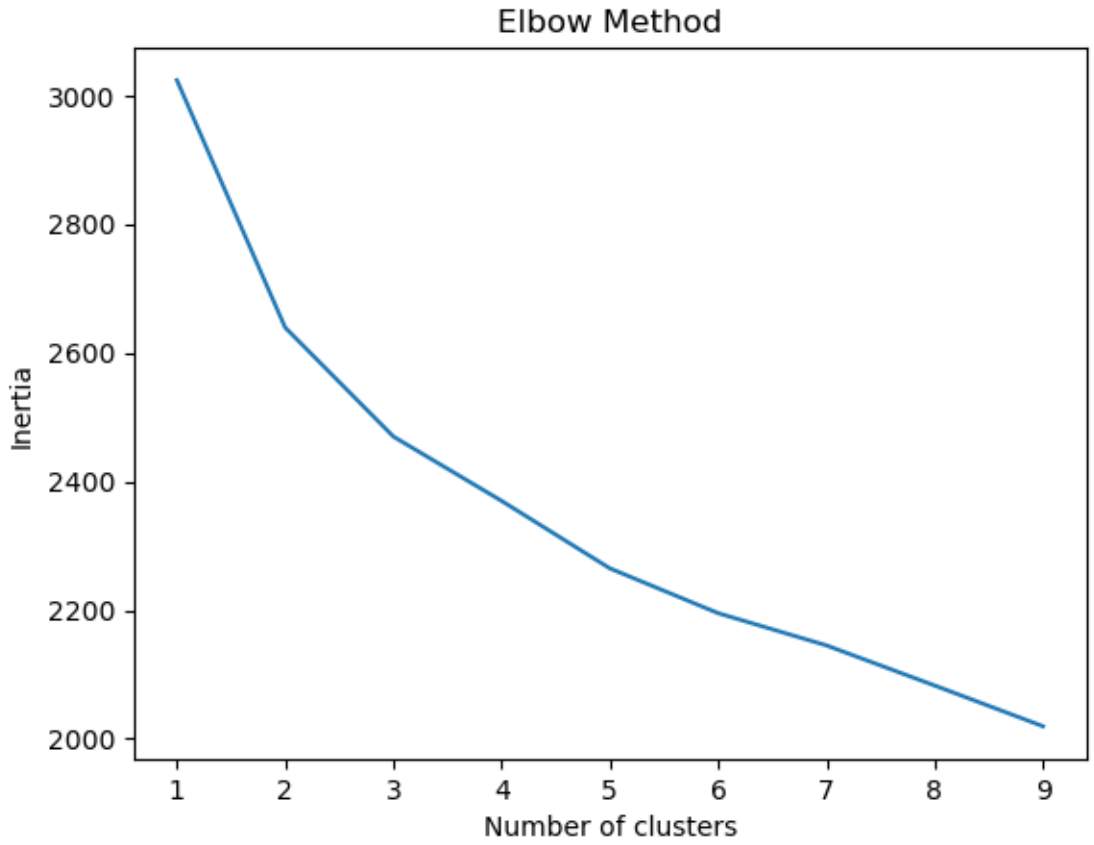
```
JavaScript: 0.0462
PowerShell: 0.0424
Go: 0.0421
                                          precision    recall  f1-score
support

                Banking/Financial Services       0.19      0.22      0.21
18
                                   Fintech       0.11      0.07      0.09
29
Internet, Telecomm or Information Services       0.33      0.22      0.26
23
              Retail and Consumer Services       0.10      0.18      0.12
11
                     Software Development       0.58      0.62      0.60
97

                                  accuracy                           0.41
178
                                 macro avg       0.26      0.26      0.26
178
                              weighted avg       0.40      0.41      0.40
178
```



0.0.7 k-means

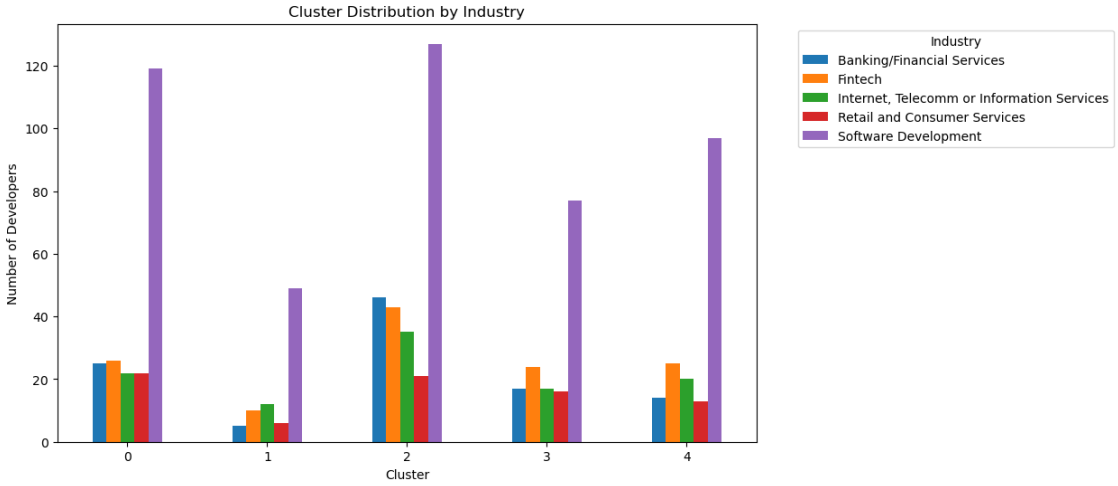
This analysis employs K-means clustering to group the UK developers based on their programming language usage patterns, derived from the Stack Overflow 2024 survey. The goal is to examine how these clusters align with the industries (Software Development, Fintech, Banking/Financial Services, Internet, Telecomm or Information Services, Retail and Consumer Service).

[44]:
```
inertia = []
for k in range(1, 10):
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(df_languages)
    inertia.append(kmeans.inertia_)

plt.plot(range(1, 10), inertia)
plt.xlabel('Number of clusters')
plt.ylabel('Inertia')
plt.title('Elbow Method')
plt.show()
```



[46]:
```
kmeans = KMeans(n_clusters=5, random_state=42)
clusters = kmeans.fit_predict(df_languages)

df_top5['Cluster'] = clusters

print("Cluster distribution by industry:")
cluster_dist = df_top5.groupby('Cluster')['Industry'].value_counts()
print(cluster_dist)

cluster_dist.unstack().plot(kind='bar', stacked=False, figsize=(10, 6))
plt.title("Cluster Distribution by Industry")
plt.xlabel("Cluster")
plt.ylabel("Number of Developers")
plt.legend(title="Industry", bbox_to_anchor=(1.05, 1), loc='upper left')
plt.xticks(rotation=0)
plt.show()
```

```
Cluster distribution by industry:
Cluster  Industry
0        Software Development                          119
         Fintech                                        26
         Banking/Financial Services                     25
         Internet, Telecomm or Information Services     22
         Retail and Consumer Services                   22
1        Software Development                           49
         Internet, Telecomm or Information Services     12
         Fintech                                        10
         Retail and Consumer Services                    6
         Banking/Financial Services                      5
2        Software Development                          127
         Banking/Financial Services                     46
         Fintech                                        43
         Internet, Telecomm or Information Services     35
         Retail and Consumer Services                   21
3        Software Development                           77
         Fintech                                        24
         Banking/Financial Services                     17
         Internet, Telecomm or Information Services     17
         Retail and Consumer Services                   16
4        Software Development                           97
         Fintech                                        25
         Internet, Telecomm or Information Services     20
         Banking/Financial Services                     14
         Retail and Consumer Services                   13
Name: count, dtype: int64
```


Cluster Distribution by Industry
Number of Developers
Cluster
0
1
2
3
4
0
20
40
60
80
100
120
Industry
Banking/Financial Services
Fintech
Internet, Telecomm or Information Services
Retail and Consumer Services
Software Development