Checkpoints on –

Mining the Insights of Stack Overflow Developer Survey

Team members:

Mohammad Imrul Jubair, CSCI 5502
Mohsena Ashraf, CSCI 5502
Cornelius Onimisi Adejoro, CSCI 5502

Project >> Overview

Stack Overflow Developer Survey (SODS)

- Survey on users to find ***how*** they learn and level up, ***which*** tools they are using, and ***what*** they want
- In May 2022 over 70,000 developers participated in SODS
  - No. of attributes: **79**
  - No of. rows: **73268**
  - Publicly available: https://insights.stackoverflow.com/survey
- Sample survey questions:

```
What is the primary operating system in which you work?
What are the primary version control systems you use?
```

Project >> Target

Mining >> Stack Overflow Developer Survey (SODS)

- Data Visualization
- Exploratory Data Analysis
- Frequent Pattern Mining
- Correlation Analysis
- Classification
- Clustering
- Hypothesis Testing

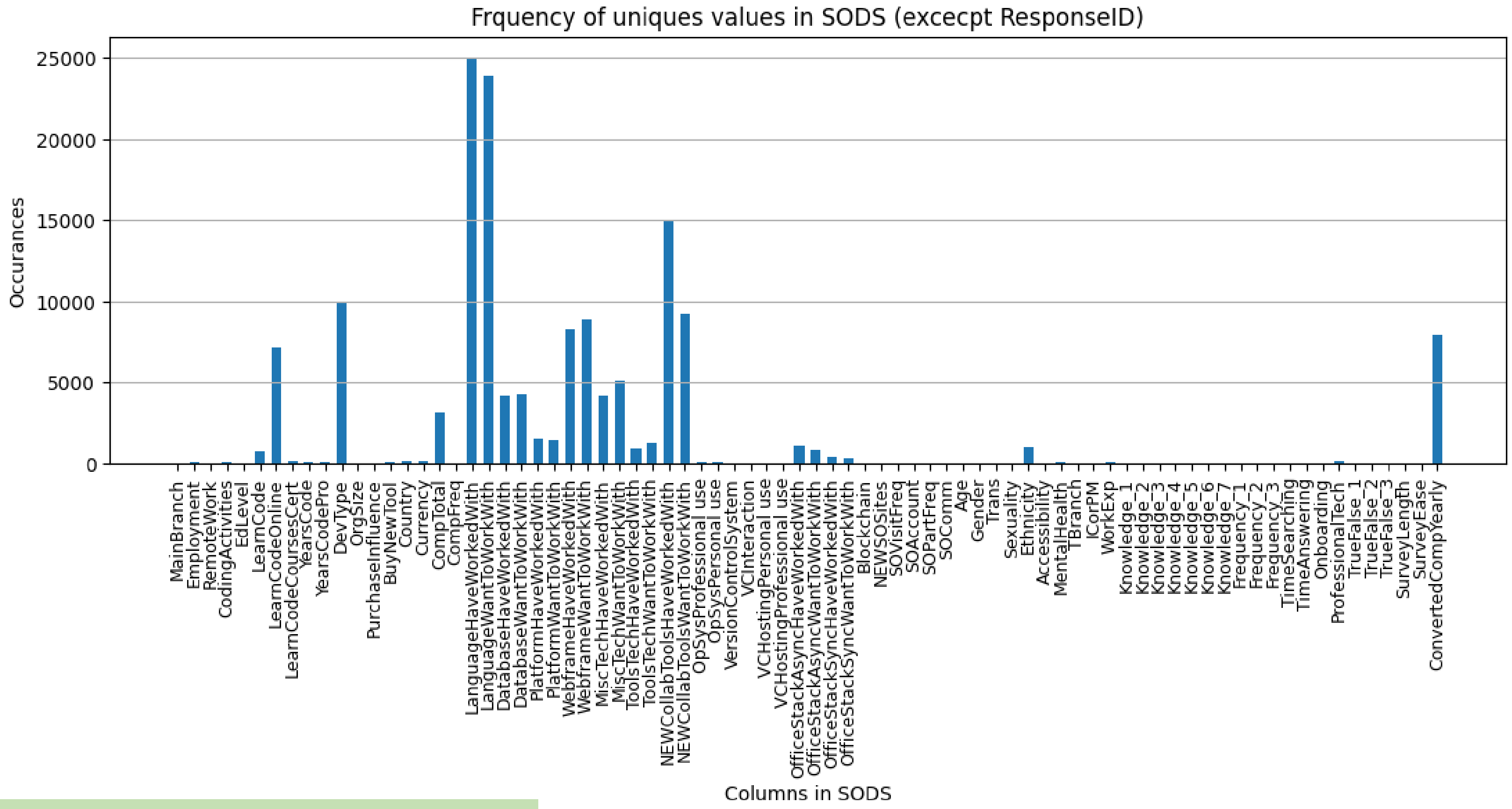
Project >> Progress

- Visualization *[mostly done]*
- Data Preprocessing
  - Cleaning *[partially done]*
  - Data Reduction *[partially done]*
  - Data Transformation *[partially done]*
- Frequent Pattern Mining *[mostly done]*
- Correlation Analysis (Chi-Square and Lift) *[mostly done]*
- Clustering *[partially done]*

Code: Python

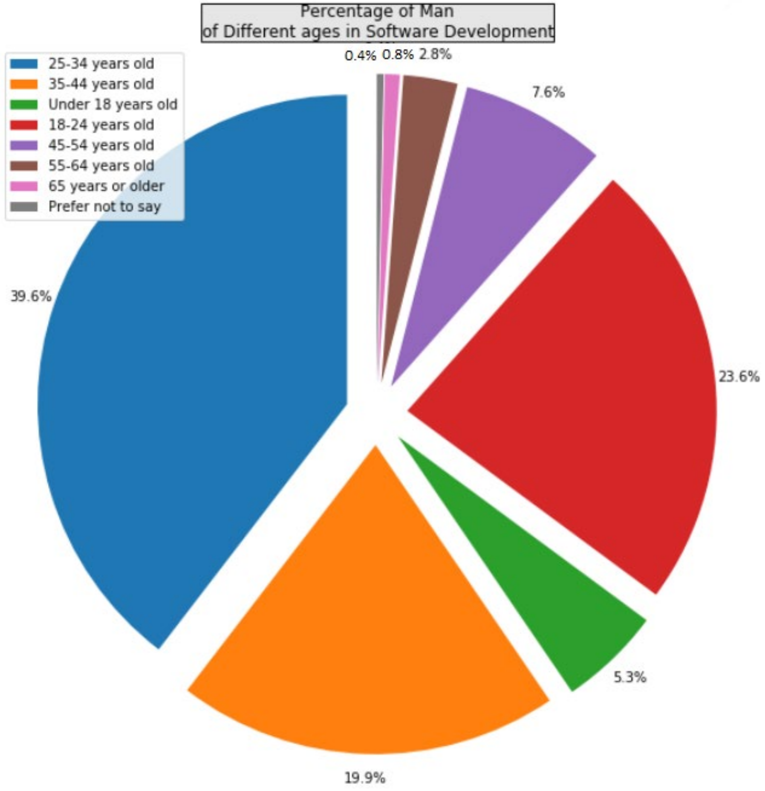
- numpy
- pandas
- scipy
- sklearn
- mlxtend
- pycountry_convert
- geopy
- geopandas
- matplotlib

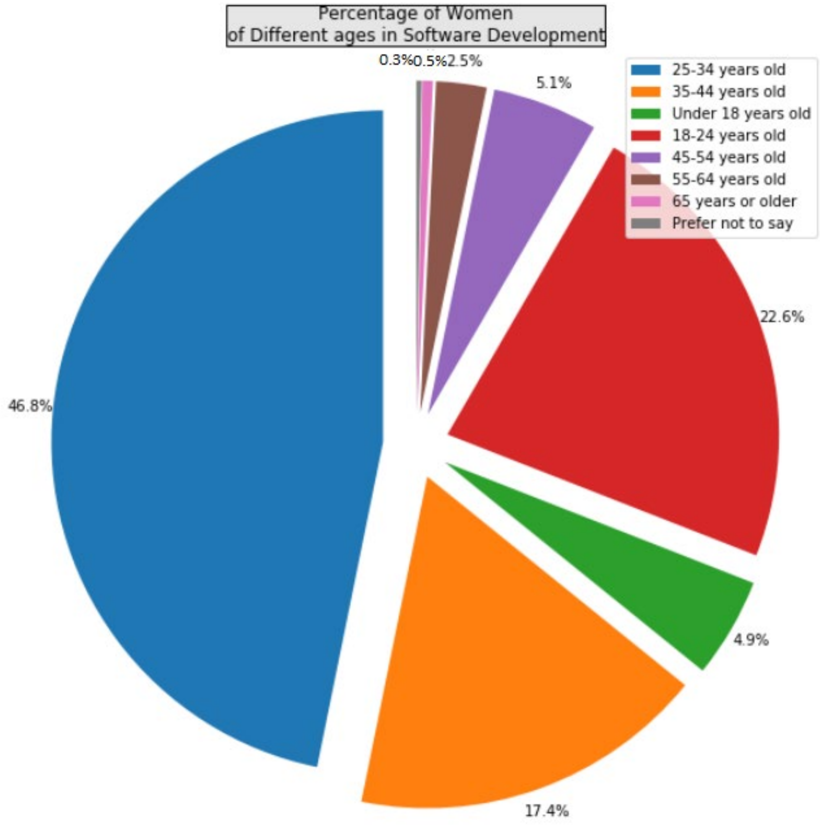
Data Visualization >> Unique Values



Insights: helpful to choose interesting attributes

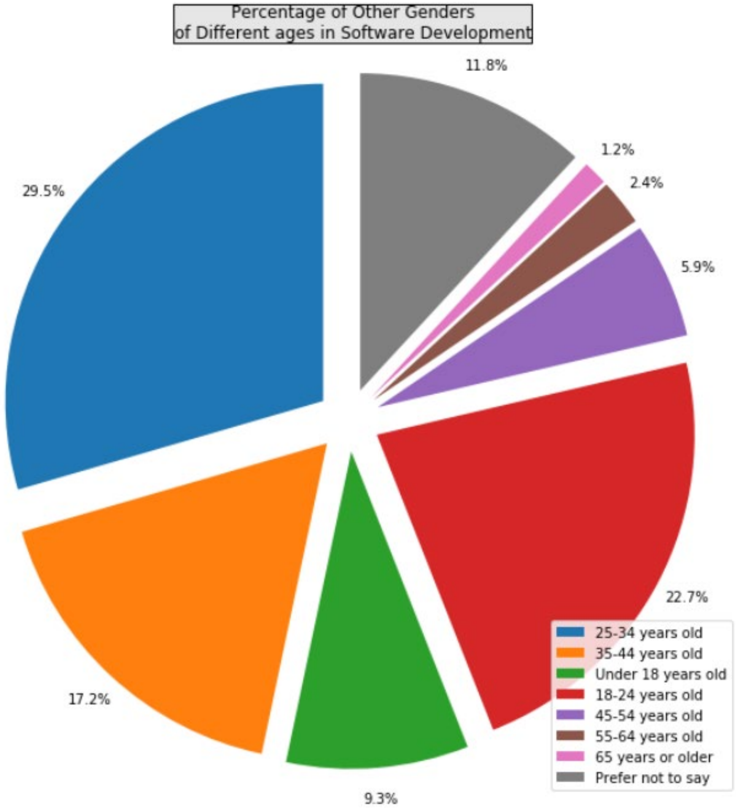
Data Visualization >> Gender : Age

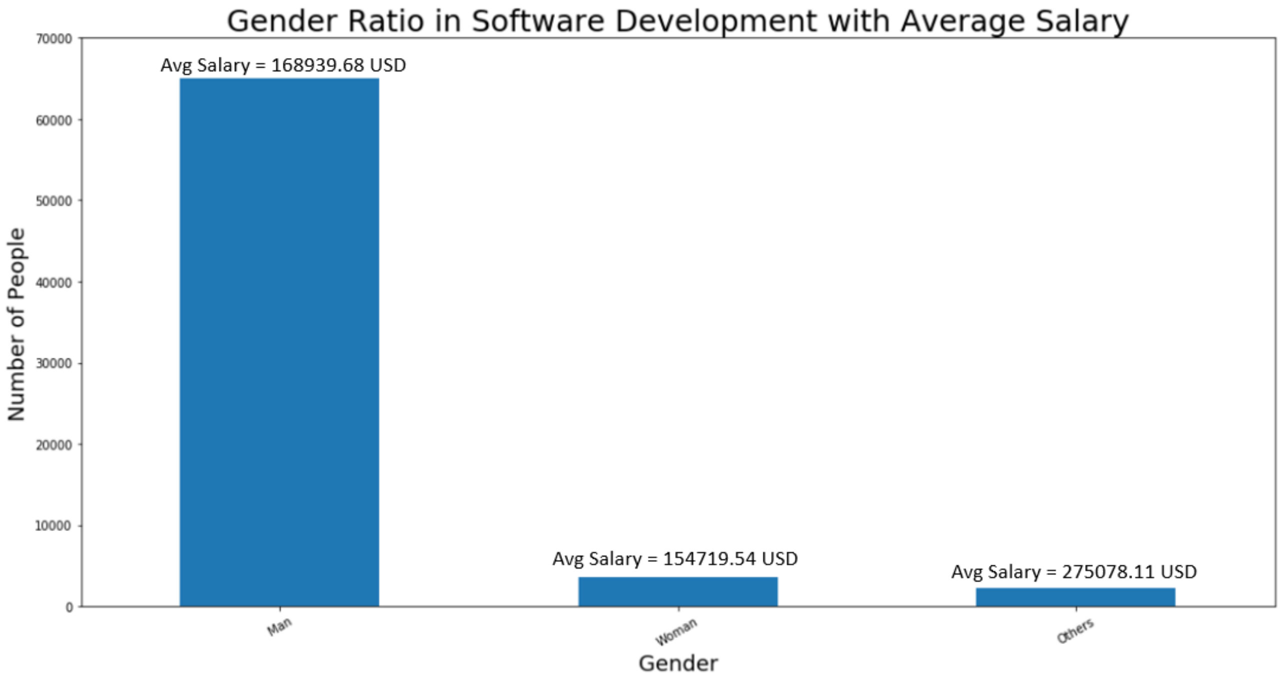

Percentage of Man
of Different ages in Software Development


Percentage of Women
of Different ages in Software Development

Insights: similar age range for male and female

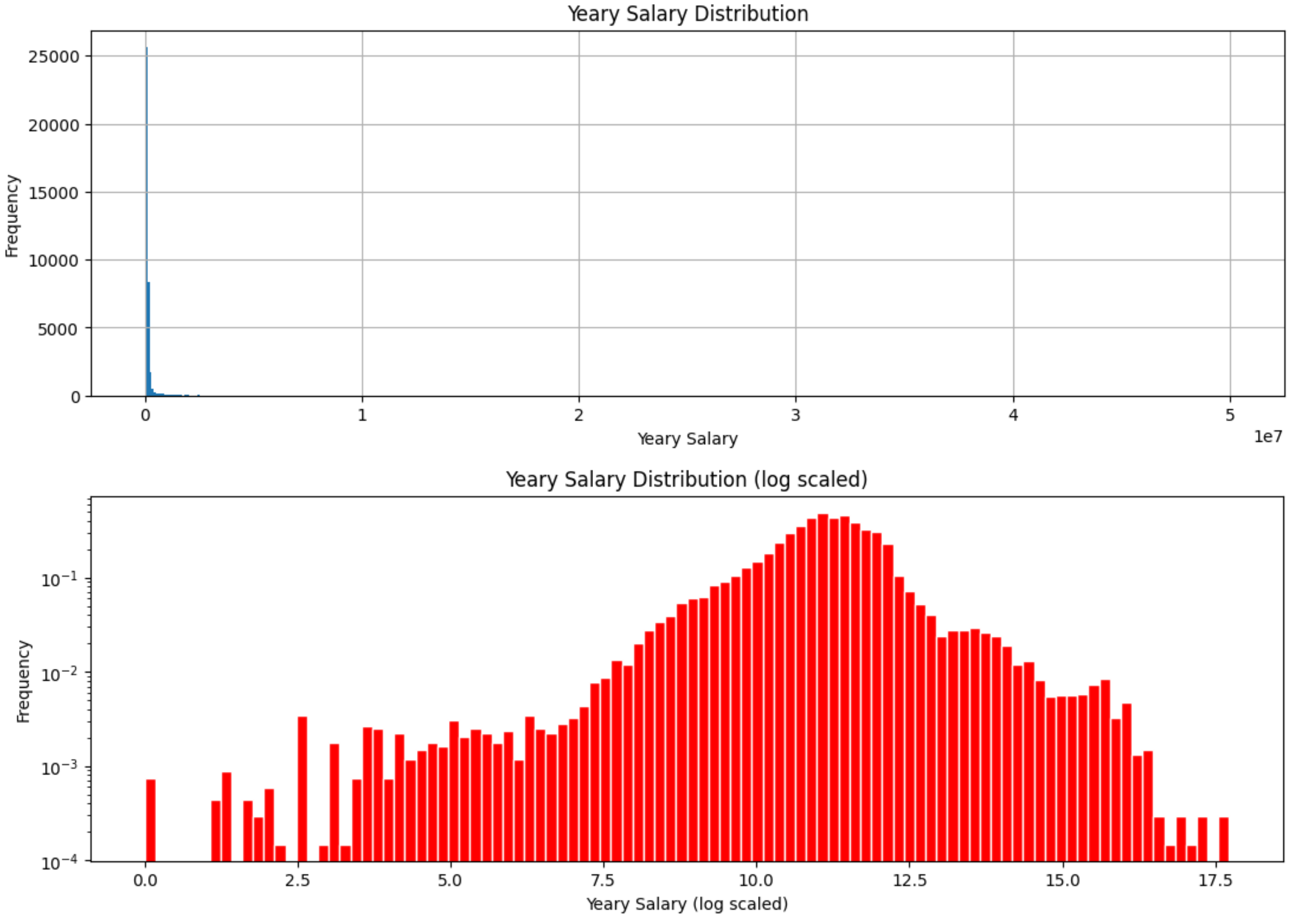
Data Visualization >> Gender : Age





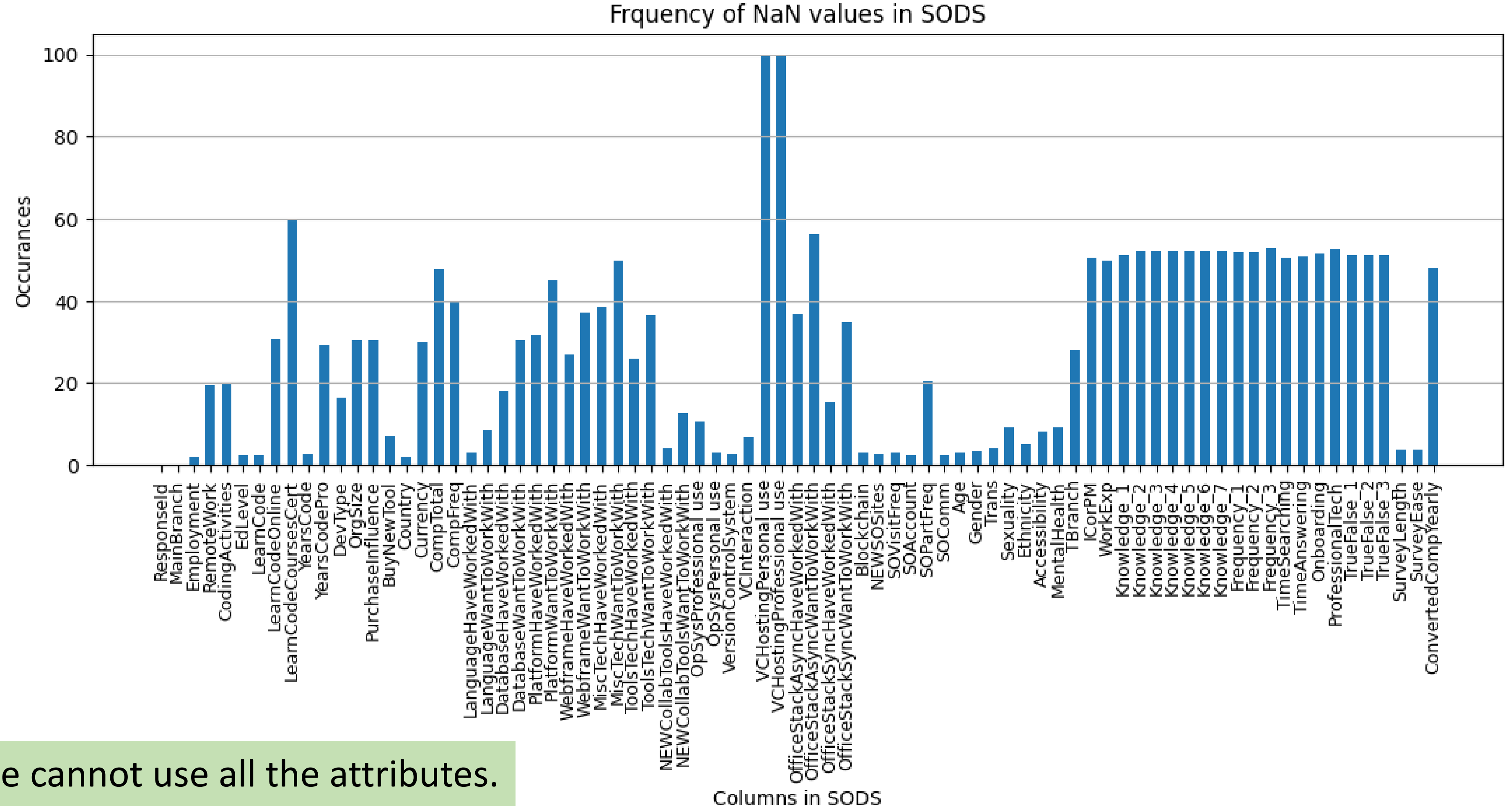
Insights: men get more salary than women

Data Visualization >> Salary



Insights: Salary might have some noise. We can find useful range.

Pre-processing >> Cleaning > NaN values



Insights: we cannot use all the attributes.

Pre-processing >> Cleaning > NaN values

Considering Less NaN:

- Usually working with attributes with less than 5% NaN values
  - ResponseId : 0.000
  - MainBranch : 0.000
  - Employment : 2.128
  - EdLevel : 2.316
  - LearnCode : 2.304
  - YearsCode : 2.644
  - Country : 2.043
  - LanguageHaveWorkedWith : 3.130
  - NEWCollabToolsHaveWorkedWith : 3.987
  - OpSysPersonal use : 3.146
  - VersionControlSystem : 2.578
  - Blockchain : 2.999
  - NEWSOSites : 2.597
  - SOVisitFreq : 3.149
  - SOAccount : 2.315
  - SOComm : 2.539
  - Age : 3.169
  - Gender : 3.296
  - Trans : 4.030
  - SurveyLength : 3.854
  - SurveyEase : 3.767

- Exceptions: for significant attributes, e.g., Salary

Pre-processing >> Reduction > Drop NaN

- Dropping NaN containing samples in a subset

Att1	Att2
A	NaN
B	C
NaN	NaN
D	E

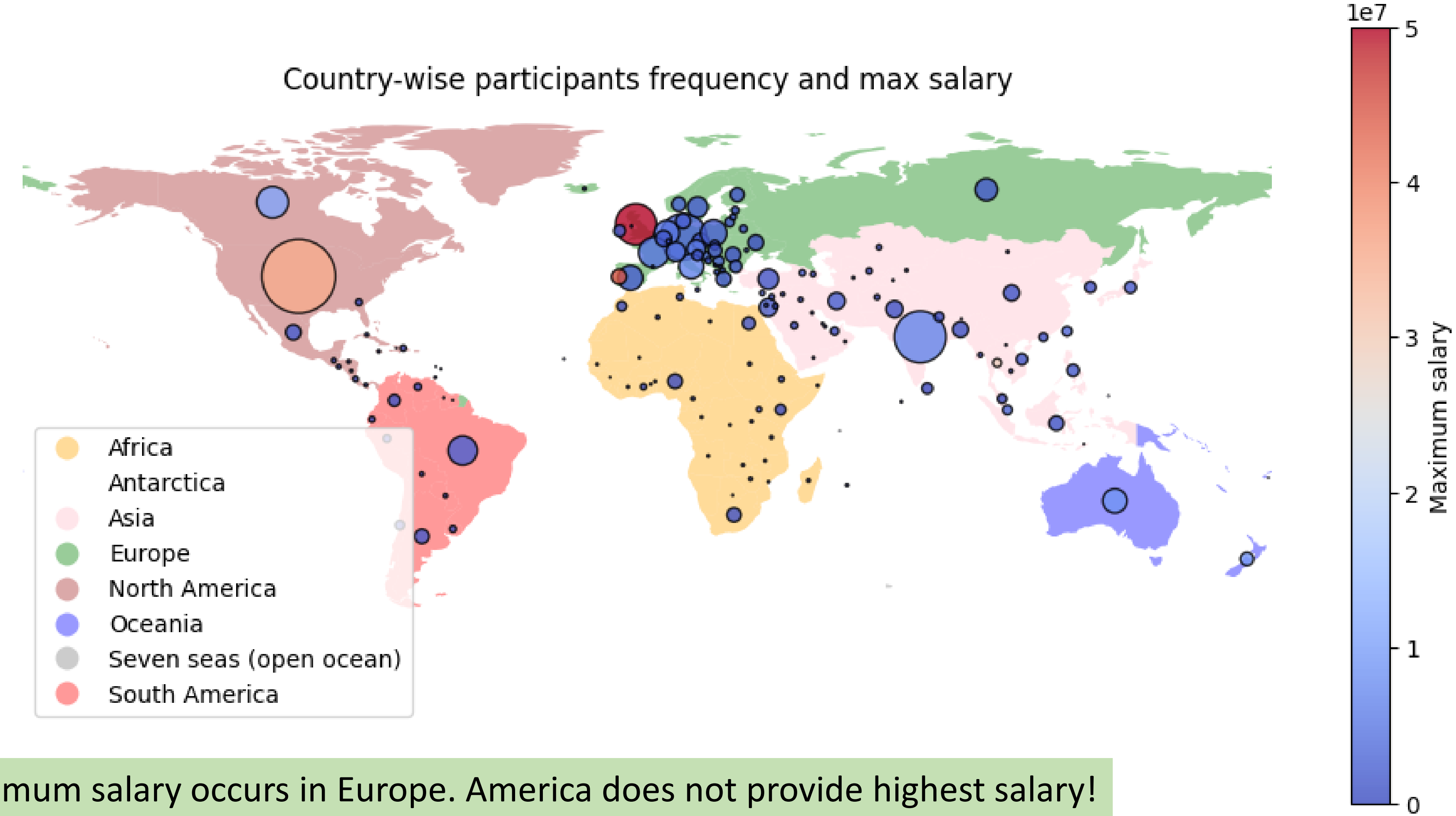
→

Att1	Att2
B	C
D	E

EDA >> Geolocation based Analysis

- Number of Countries: 181
    - Which countries has more participants?
    - Developers from which countries get more salaries?
- *Challenges:*
    - Not all the countries were in common format
        - e.g., The former Yugoslav Republic of Macedonia → North Macedonia
    - Latitude Longitude of the countries
    - Visualizing frequencies
        - Multimodal

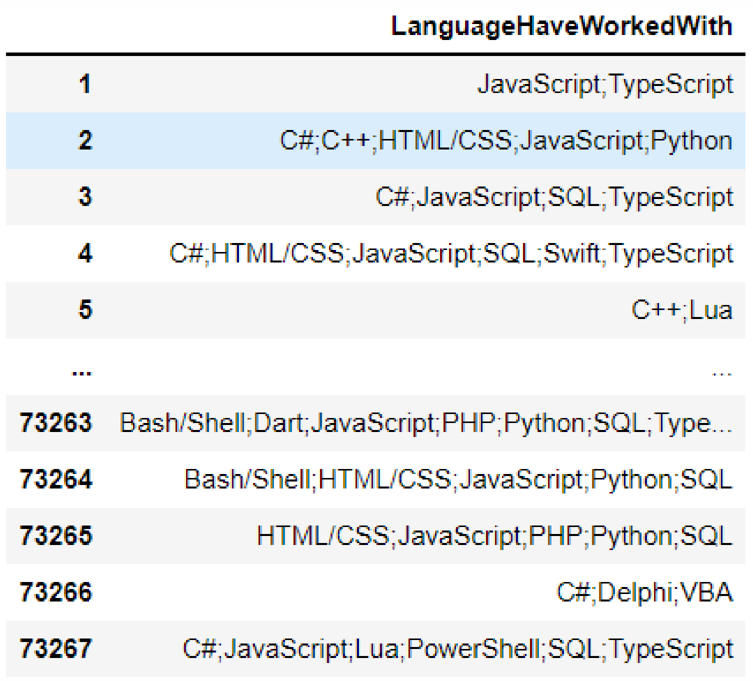
EDA >> Geolocation based Analysis



Insights: Maximum salary occurs in Europe. America does not provide highest salary!

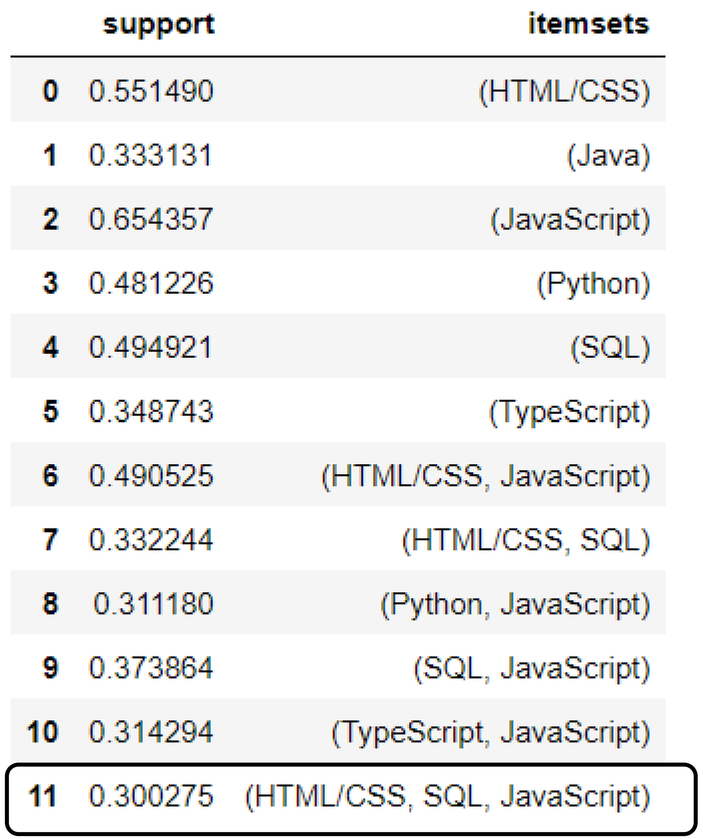
FPA >> Apriori

- Finding the *most frequently used programming language*



	LanguageHaveWorkedWith
1	JavaScript;TypeScript
2	C#;C++;HTML/CSS;JavaScript;Python
3	C#;JavaScript;SQL;TypeScript
4	C#;HTML/CSS;JavaScript;SQL;Swift;TypeScript
5	C++;Lua
...	...
73263	Bash/Shell;Dart;JavaScript;PHP;Python;SQL;Type...
73264	Bash/Shell;HTML/CSS;JavaScript;Python;SQL
73265	HTML/CSS;JavaScript;PHP;Python;SQL
73266	C#;Delphi;VBA
73267	C#;JavaScript;Lua;PowerShell;SQL;TypeScript

min_supp = 0.3

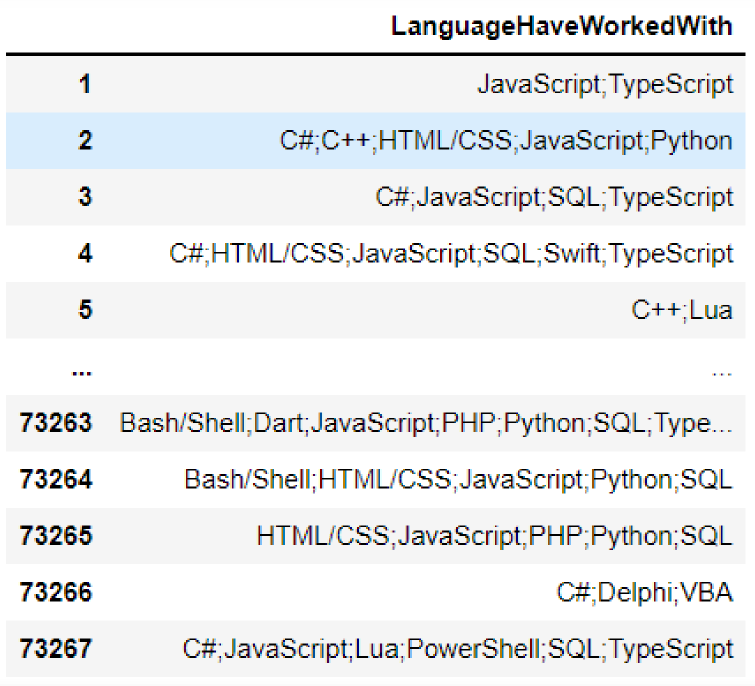


	support	itemsets
0	0.551490	(HTML/CSS)
1	0.333131	(Java)
2	0.654357	(JavaScript)
3	0.481226	(Python)
4	0.494921	(SQL)
5	0.348743	(TypeScript)
6	0.490525	(HTML/CSS, JavaScript)
7	0.332244	(HTML/CSS, SQL)
8	0.311180	(Python, JavaScript)
9	0.373864	(SQL, JavaScript)
10	0.314294	(TypeScript, JavaScript)
11	0.300275	(HTML/CSS, SQL, JavaScript)

Insights: frequently used PL = {HTML/CSS, SQL, JavaScript} (k=3)

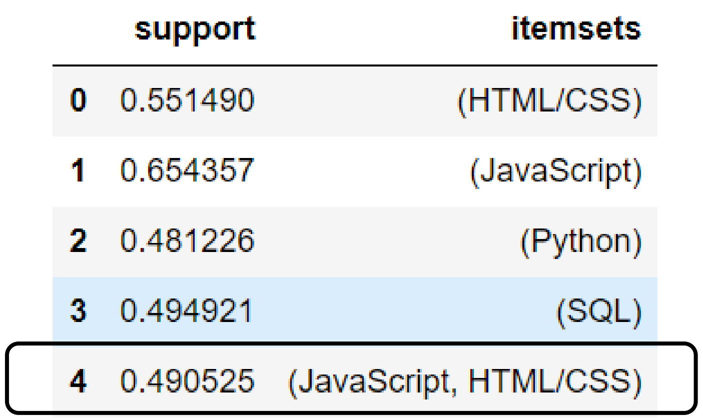
FPA >> Apriori

- Finding the *most frequently used programming language*



	LanguageHaveWorkedWith
1	JavaScript;TypeScript
2	C#;C++;HTML/CSS;JavaScript;Python
3	C#;JavaScript;SQL;TypeScript
4	C#;HTML/CSS;JavaScript;SQL;Swift;TypeScript
5	C++;Lua
...	...
73263	Bash/Shell;Dart;JavaScript;PHP;Python;SQL;Type...
73264	Bash/Shell;HTML/CSS;JavaScript;Python;SQL
73265	HTML/CSS;JavaScript;PHP;Python;SQL
73266	C#;Delphi;VBA
73267	C#;JavaScript;Lua;PowerShell;SQL;TypeScript

min_supp = 0.4

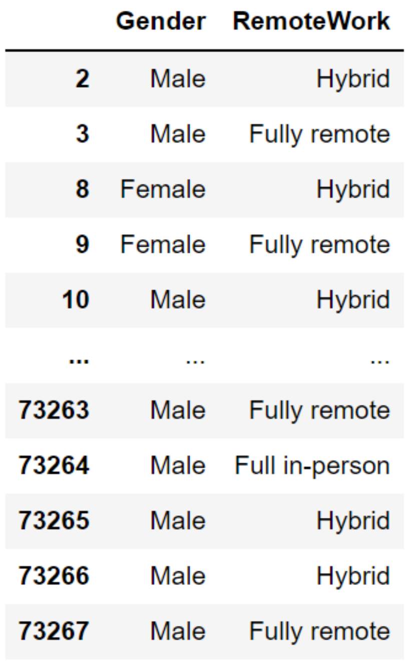


	support	itemsets
0	0.551490	(HTML/CSS)
1	0.654357	(JavaScript)
2	0.481226	(Python)
3	0.494921	(SQL)
4	0.490525	(JavaScript, HTML/CSS)

Insights: frequently used PL = {HTML/CSS, JavaScript} (k=2)

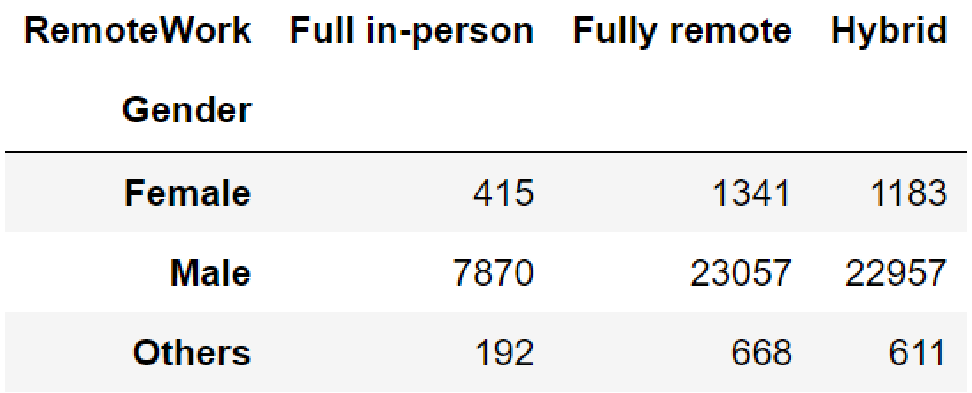
Correlation Analysis >> Lift Measurement

- Remote Work:
  - {Full in-person, fully remote, hybrid}
- Gender:
  - {Male, Female, Others}
- *Challenge*
  - Gender attribute is noisy (multiple answers)



	Gender	RemoteWork
2	Male	Hybrid
3	Male	Fully remote
8	Female	Hybrid
9	Female	Fully remote
10	Male	Hybrid
...	...	...
73263	Male	Fully remote
73264	Male	Full in-person
73265	Male	Hybrid
73266	Male	Hybrid
73267	Male	Fully remote

Correlation Analysis >> Lift Measurement

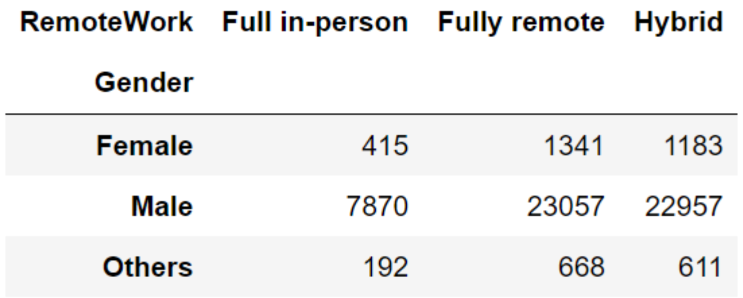


RemoteWork	Full in-person	Fully remote	Hybrid
Gender			
Female	415	1341	1183
Male	7870	23057	22957
Others	192	668	611

```
lift( 0 , 0 ) = 0.9710 [-ve corr]
lift( 0 , 1 ) = 1.0611 [+ve corr]
lift( 0 , 2 ) = 0.9480 [-ve corr]
lift( 1 , 0 ) = 1.0044 [+ve corr]
lift( 1 , 1 ) = 0.9951 [-ve corr]
lift( 1 , 2 ) = 1.0034 [+ve corr]
lift( 2 , 0 ) = 0.8976 [-ve corr]
lift( 2 , 1 ) = 1.0561 [+ve corr]
lift( 2 , 2 ) = 0.9783 [-ve corr]
```

Insights: lift (*Female*, *Fully remote*) = positively correlated
Negatively correlated for (*Female, in-person) and* (*Female, hybrid)*

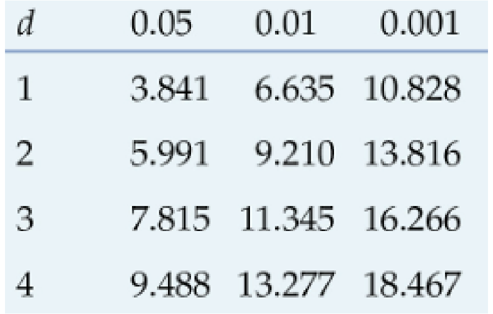
Correlation Analysis >> Chi-square Test



RemoteWork	Full in-person	Fully remote	Hybrid
Gender			
Female	415	1341	1183
Male	7870	23057	22957
Others	192	668	611

- **X^2 value =** 13.950, **DoF** = 4, **Significance Level, a** = 0.05
- **Correlated, as X^2 > 9.488**
- Expected:

427.383	1263.748	1247.867
7835.706	23169.731	22878.561
213.909	632.519	624.570

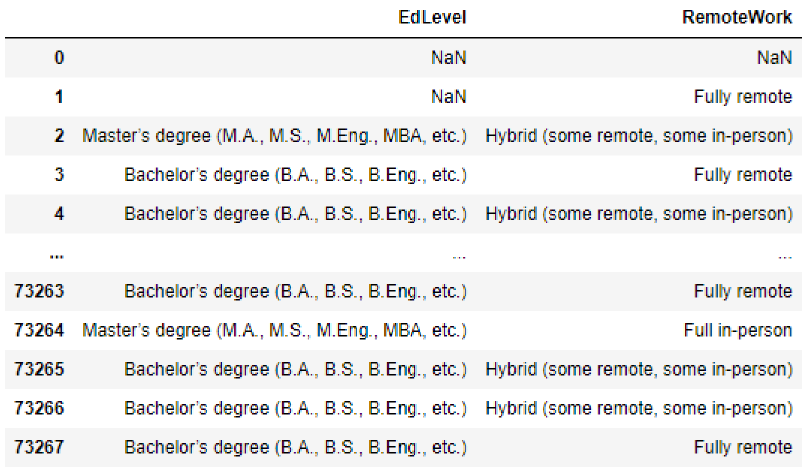


d	0.05	0.01	0.001
1	3.841	6.635	10.828
2	5.991	9.210	13.816
3	7.815	11.345	16.266
4	9.488	13.277	18.467

Insights: [o11 < e11; o12 > e12; o13 < e13] Female persons are more likely to work remotely, rather than working in person or hybrid.

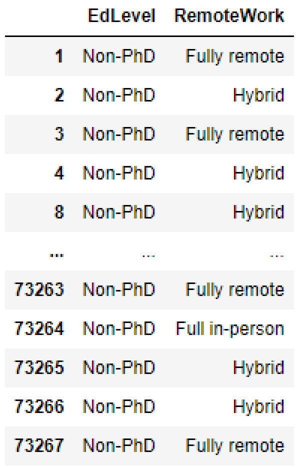
Correlation Analysis >> Chi-square Test

- Remote Work:
  - {Full in-person, fully remote, hybrid}
- Education Level:
  - {PhD, Non-PhD}
- *Challenge:*
  - Education Level has many unique values



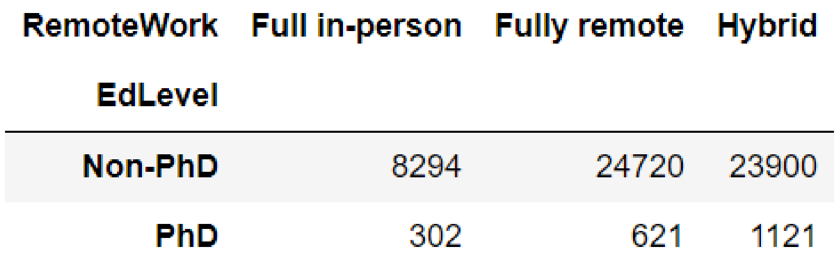
	EdLevel	RemoteWork
0	NaN	NaN
1	NaN	Fully remote
2	Master's degree (M.A., M.S., M.Eng., MBA, etc.)	Hybrid (some remote, some in-person)
3	Bachelor's degree (B.A., B.S., B.Eng., etc.)	Fully remote
4	Bachelor's degree (B.A., B.S., B.Eng., etc.)	Hybrid (some remote, some in-person)
...	...	...
73263	Bachelor's degree (B.A., B.S., B.Eng., etc.)	Fully remote
73264	Master's degree (M.A., M.S., M.Eng., MBA, etc.)	Full in-person
73265	Bachelor's degree (B.A., B.S., B.Eng., etc.)	Hybrid (some remote, some in-person)
73266	Bachelor's degree (B.A., B.S., B.Eng., etc.)	Hybrid (some remote, some in-person)
73267	Bachelor's degree (B.A., B.S., B.Eng., etc.)	Fully remote

↓



	EdLevel	RemoteWork
1	Non-PhD	Fully remote
2	Non-PhD	Hybrid
3	Non-PhD	Fully remote
4	Non-PhD	Hybrid
8	Non-PhD	Hybrid
...	...	...
73263	Non-PhD	Fully remote
73264	Non-PhD	Full in-person
73265	Non-PhD	Hybrid
73266	Non-PhD	Hybrid
73267	Non-PhD	Fully remote

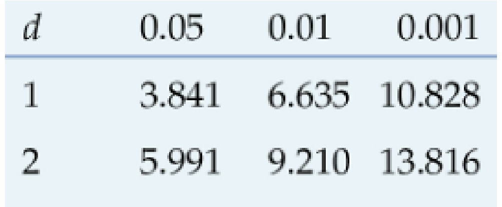
Correlation Analysis >> Chi-square Test



RemoteWork	Full in-person	Fully remote	Hybrid
EdLevel			
Non-PhD	8294	24720	23900
PhD	302	621	1121

- **X^2 value =** 155.039, **DoF** = 2, **Significance Level, a** = 0.05
- **Correlated, as X^2 > 5.991**
- **Expected:**

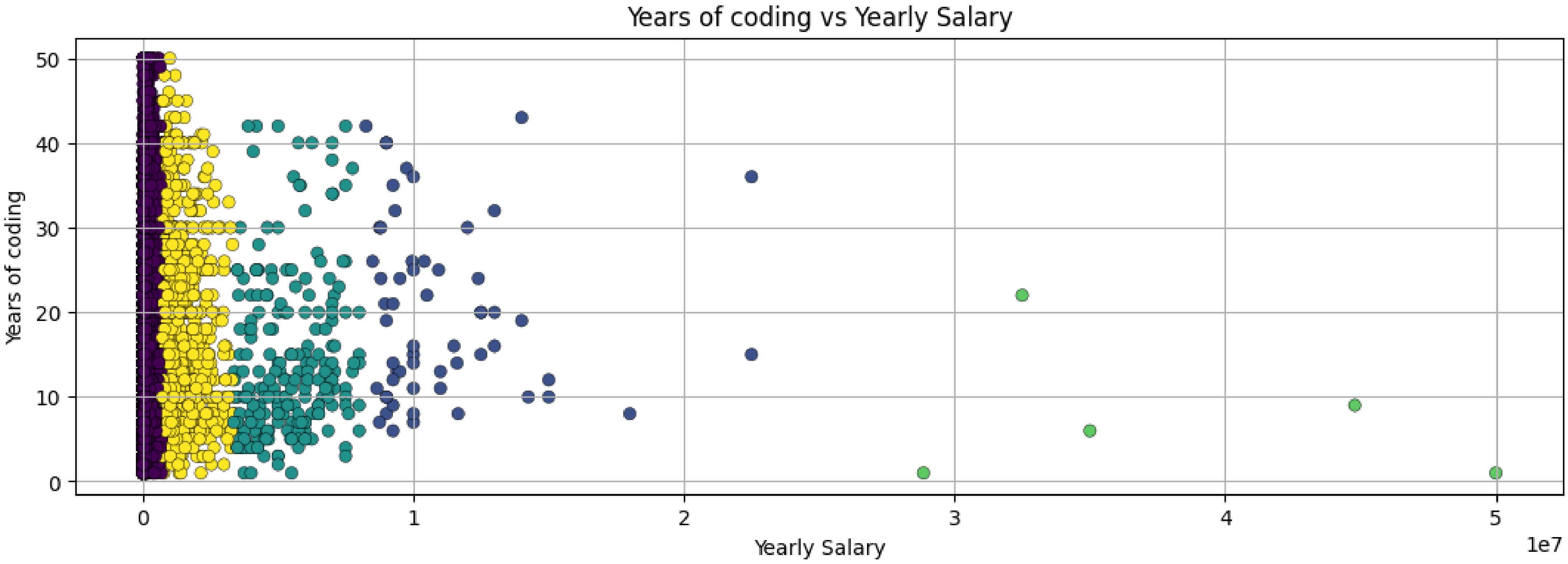
8297.987	24462.459	24153.5534
298.012	878.540	867.446



d	0.05	0.01	0.001
1	3.841	6.635	10.828
2	5.991	9.210	13.816

Insights: [o21 > e21; o22 < e22; o23 > e23] PhD persons are more likely to work in person or hybrid, rather than working remotely.

K-means Clustering >>



Insights: can be used to determine noise

Future Work >>

- Completing “partially done tasks”
- Present most interesting mining outcomes.
- Hypothesis Testing:
    - e.g., given the same education level, skills and experience do women get the same salary as men employee?
        - Naïve Bayesian
        - Decision Tree
- Classification:
    - e.g., given education level, experience in language and tools
        - What salary one should expect?
            - Low, mid or high?
- Clustering:
    - K-medoids Clustering for Categorical Values

Thank You

Any Comments?