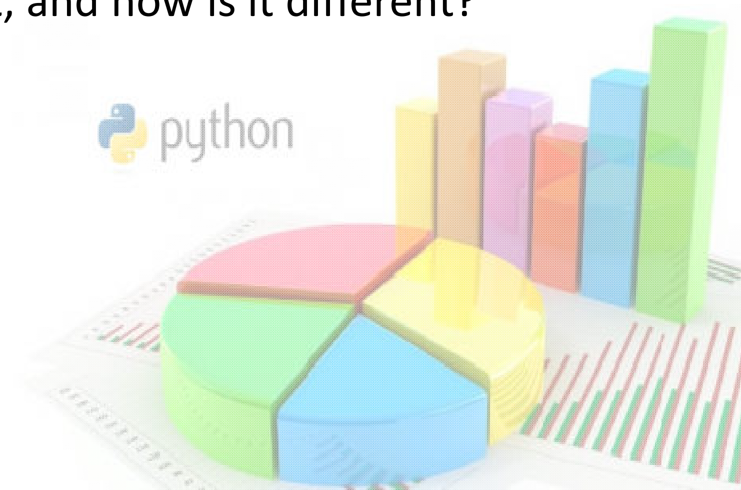
Ramsey A.

Data Science Employment Status

- Major Research Question:
  - **What is the current status of Data Science?**

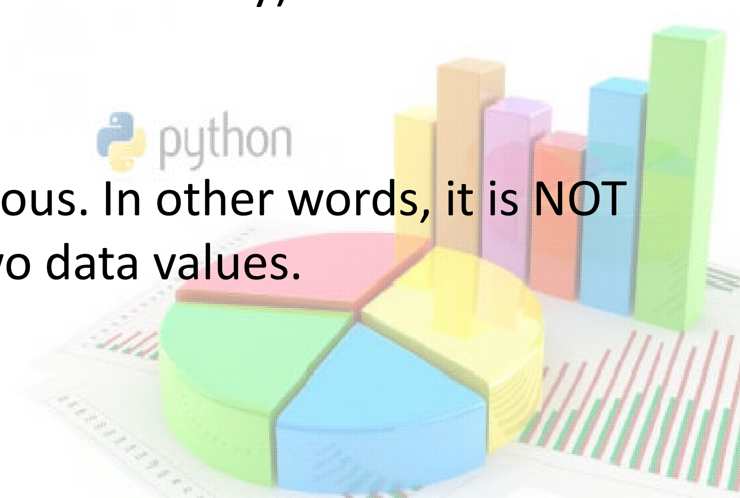
- Other Questions:
  - What are the key educational backgrounds of the current Data Scientists and Data Analysts?
  - What are the key programming languages ? Do they really increase the chances of employment?
  - Who switch career into Data Science or Data Analysis? Why?
  - What are the chances of female data scientist/analyst to find a job?
  - What are the key features of Canada’s D.S market, and how is it different?

- Open Questions:
  - **Am I learning the right thing?**
  - **Is Metro Bluffing?**



The Statistical Methodology:

- **Logistic regression (GLM) Method**
  - The binary logistic model is used to estimate the probability of a binary response (Dependent Variable) based on one or more predictor (or independent) variables (features).
  - It allows one to say that the presence of a risk factor increases the odds of a given outcome by a specific factor. The model itself simply models probability of output in terms of input.
  - The normal (*z*) distribution is a continuous distribution, which means that between any two data values we could (at least in theory) find another data value.
  - Binomial distribution is discrete, not continuous. In other words, it is NOT possible to find a data value between any two data values.



Project Data

- Kaggle’s survey (2017-2018) to establish a comprehensive view of the state of data science and machine learning. The data set contains **16,000 responses** and covering who is working with data, what’s happening at the cutting edge of machine learning across industries, and how new data scientists can best break into the field.
- Data subset was created including the following variables:
  - ‘Gender', 'Country', 'Age', 'Employment', 'Student_Status', 'Code_Writer', 'Career_Switcher', 'Current_Job_Title', 'Language_Recommendation', 'Time_Spent_Studying', 'Education', 'Field_of_Education'
- A Canadian Data set was sliced to compare Canada to the International Market.

Python Methodology

- **<u>Stage One: Data Cleaning & Manipulation</u>**

  Libraries & Packages used:

  import numpy as np   import pandas as pd   import os

  import matplotlib.pyplot as plt   import statsmodels.api as sm

  Import statsmodels.formula.api as smf   import seaborn as sns

- Browsing & Selecting Relevant Variables
- Data Cleaning and Conversion : to suit the statistical methodology

1. **<u>Binning</u>**:

   G_bins = [0,1,2,3,4]

2. **<u>Grouping & Labeling</u>**:

   G_labels = {"Non-binary, genderqueer, or gender non-conforming": 0, "A different identity": 1, "Female": 2, "Male": 3}

3. **<u>Re-Categorizing & Coding</u>**:

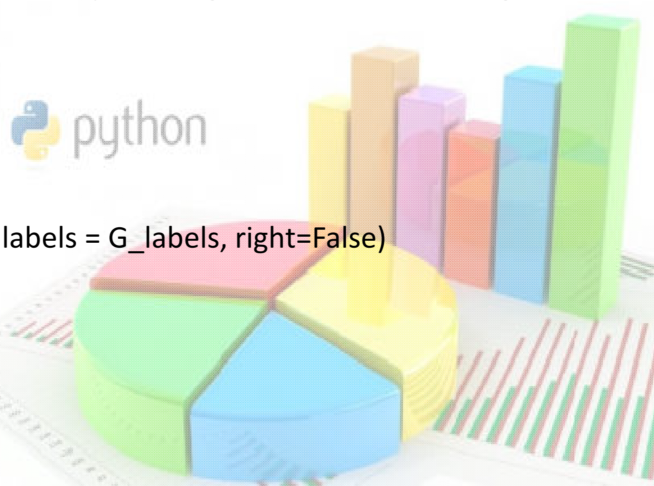
   Data_Scientist['Gender_Cat'] = coding(Data_Scientist['Gender'], {"Non-binary, genderqueer, or gender non-conforming": 0, "A different identity": 1, "Female": 2, "Male": 3})

4. **<u>Cleaning & Manipulation</u>**:

   Data_Scientist['Gender_Cat'] = Data_Scientist['Gender_Cat'].fillna(0)

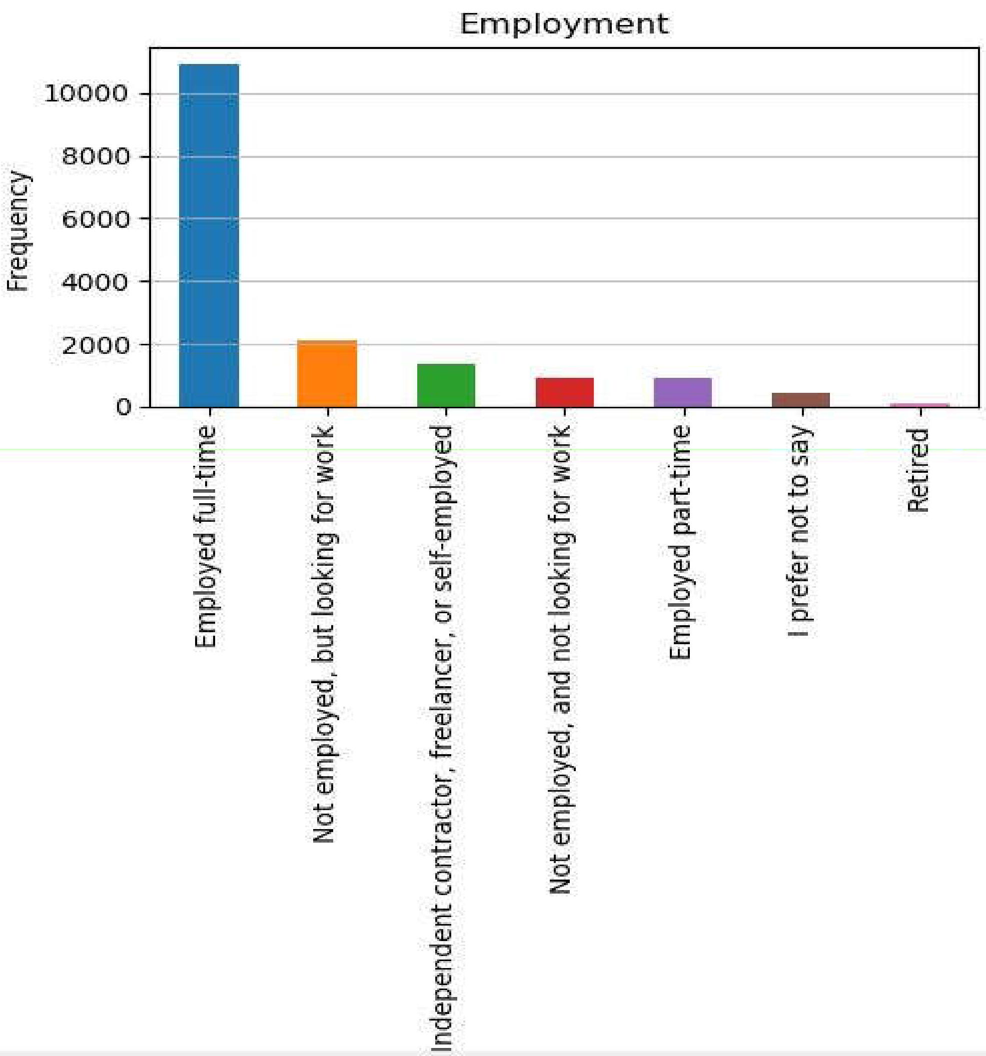
5. **<u>Reframing & restructuring</u>**:

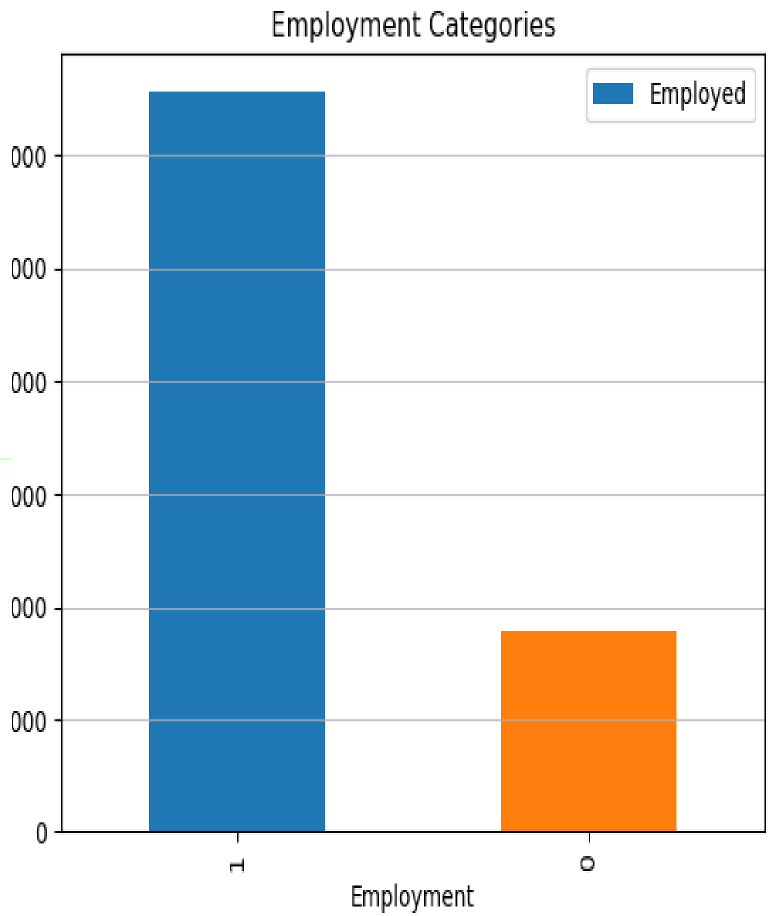
   Data_Scientist['Gender_Labeled'] = pd.cut(Data_Scientist.Gender_Cat, G_bins, labels = G_labels, right=False)

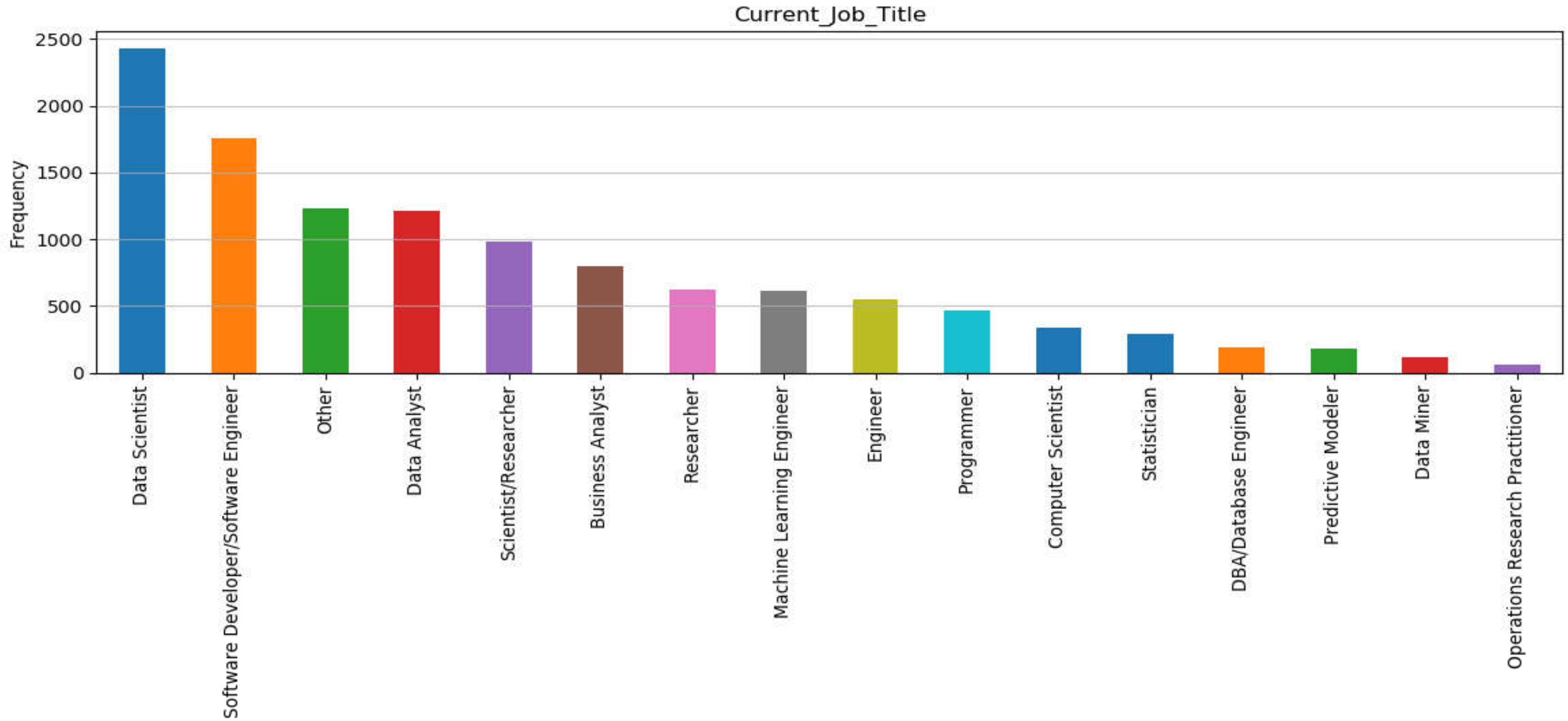


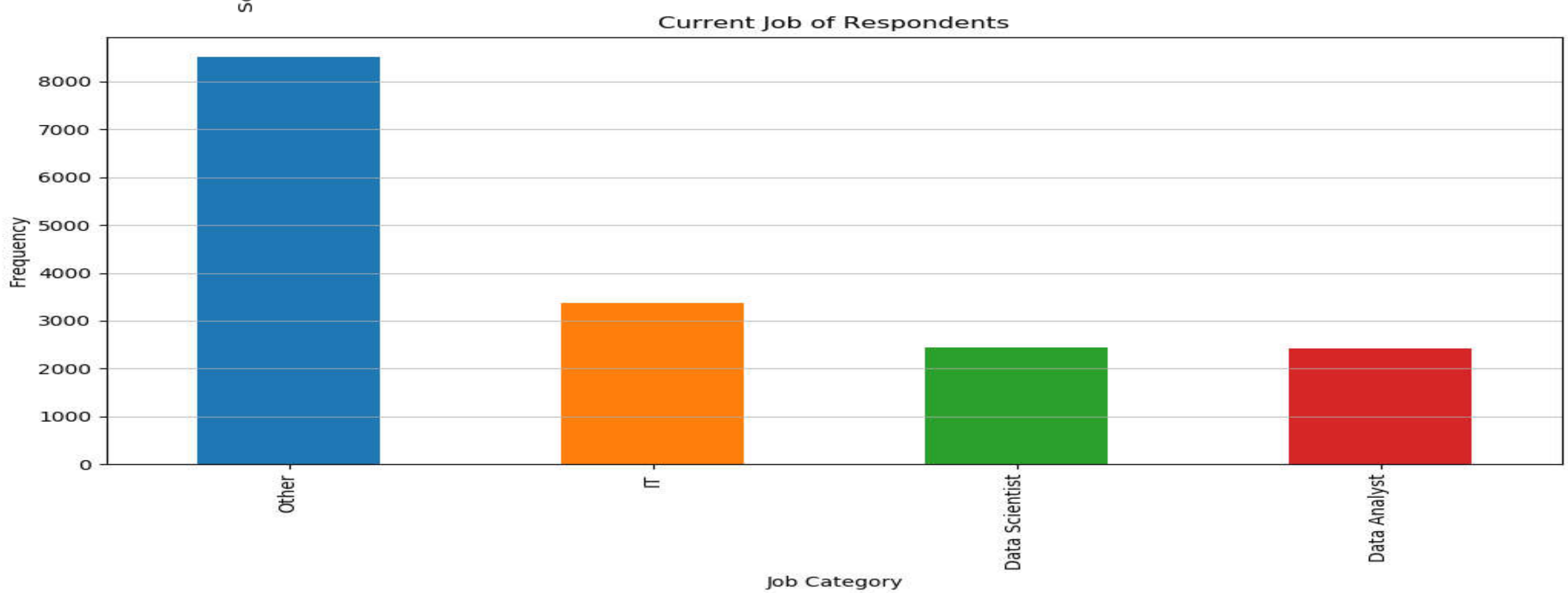
- **<u>Stage 2: plotting the variables:</u>** <u>Descriptive Statistics</u>

(Examples on Original Data Vs. Coded/Categorized)

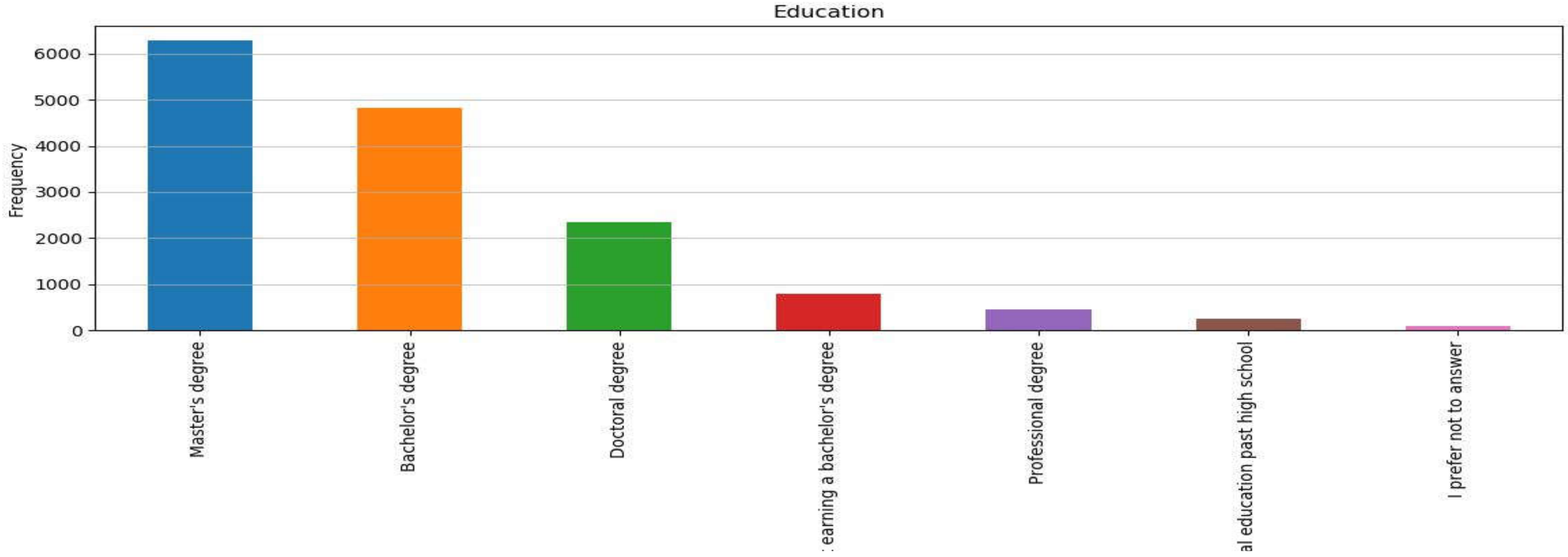


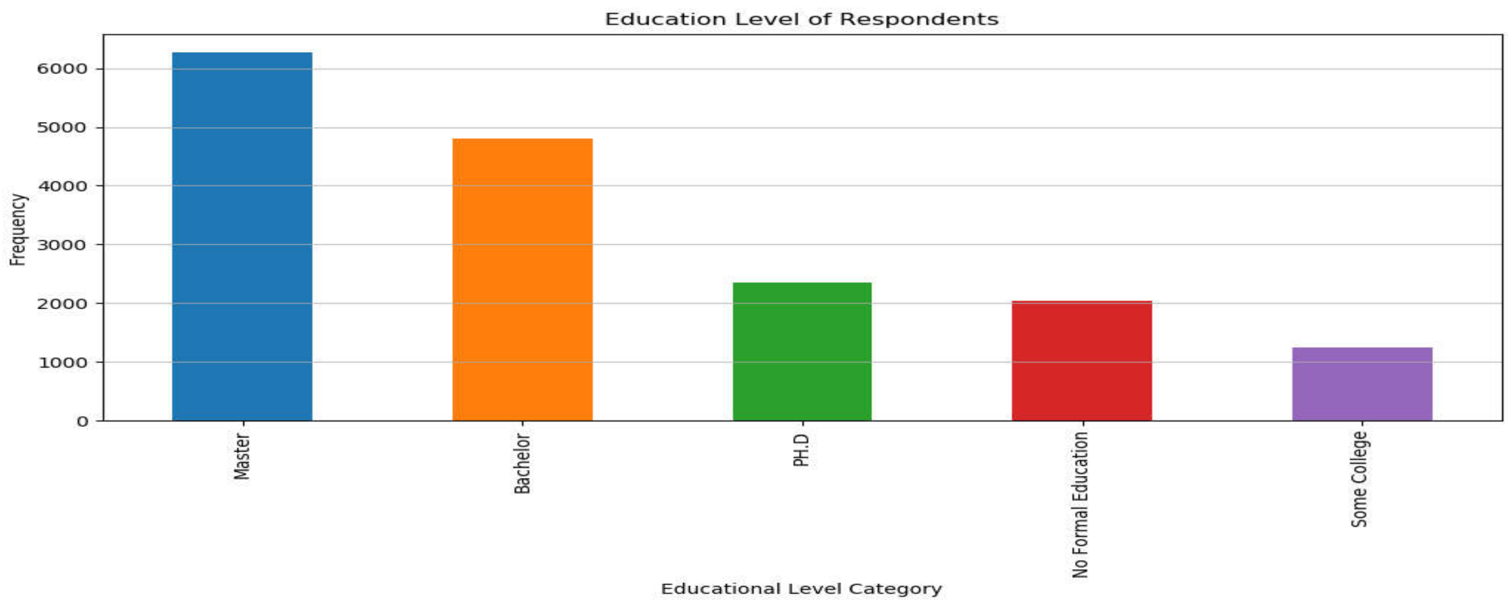



Current_Job_Title
Frequency
0
500
1000
1500
2000
2500
Data Scientist
Software Developer/Software Engineer
Other
Data Analyst
Scientist/Researcher
Business Analyst
Researcher
Machine Learning Engineer
Engineer
Programmer
Computer Scientist
Statistician
DBA/Database Engineer
Predictive Modeler
Data Miner
Operations Research Practitioner

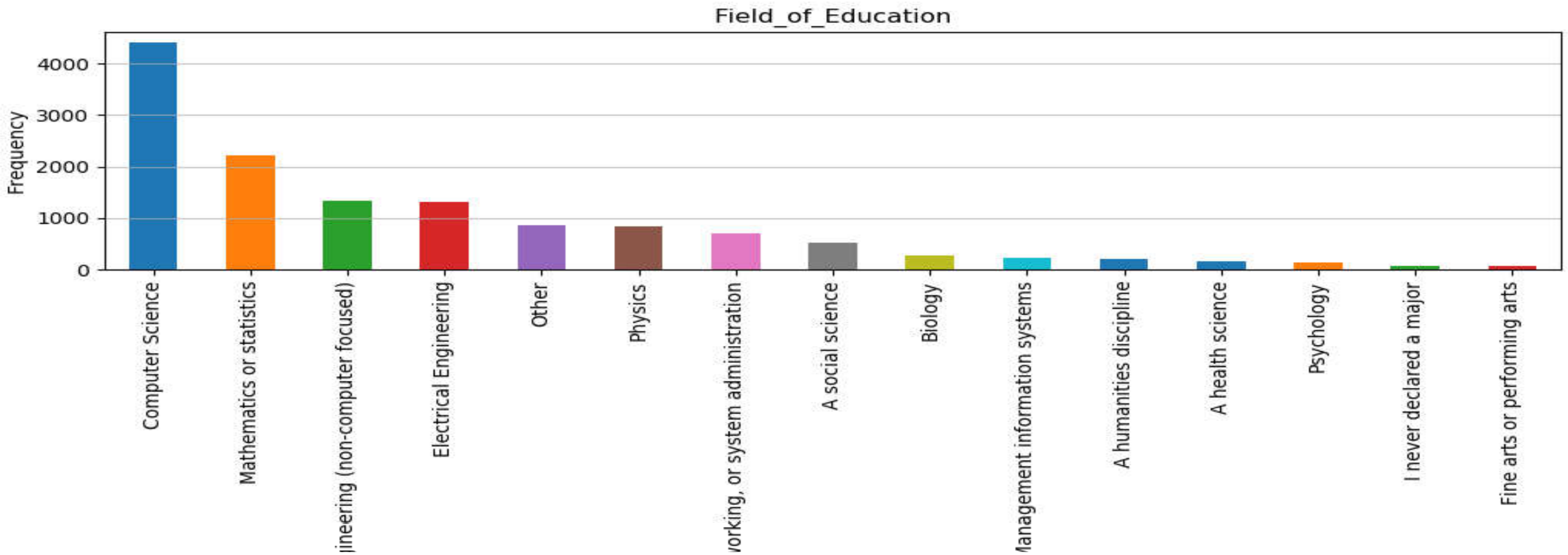

Current Job of Respondents
Frequency
0
1000
2000
3000
4000
5000
6000
7000
8000
Other
IT
Data Scientist
Data Analyst
Job Category

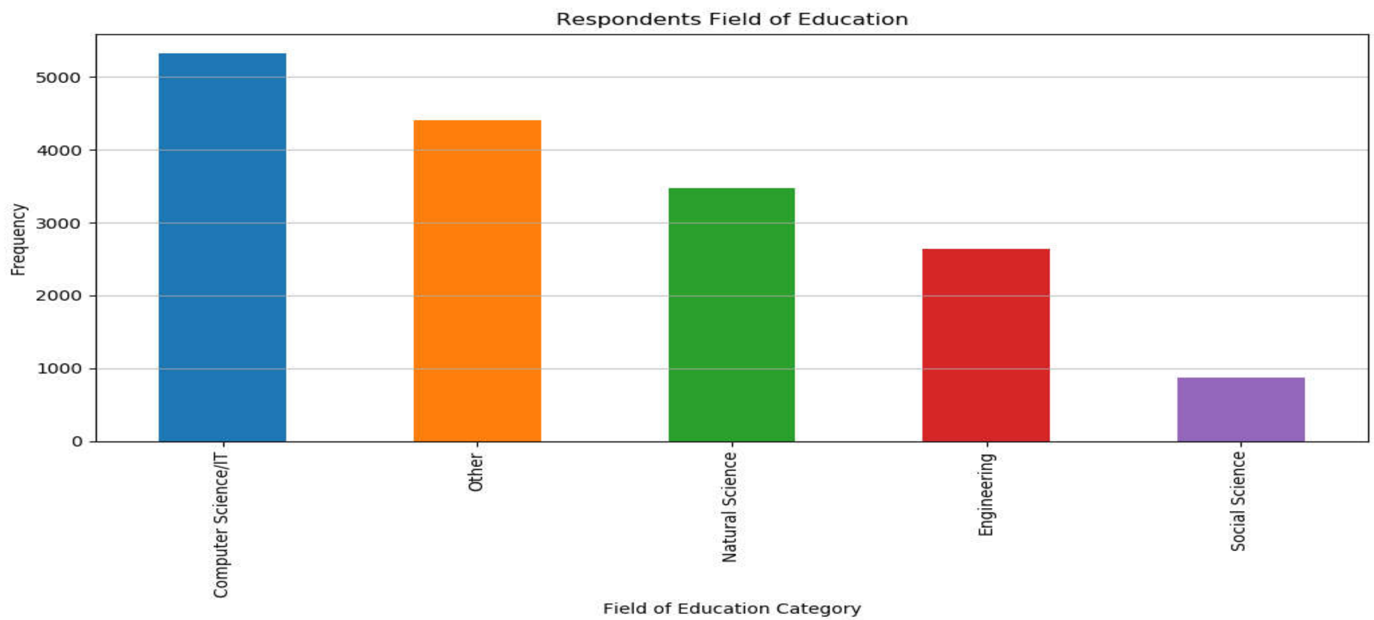
- Stage 2: plotting the variables



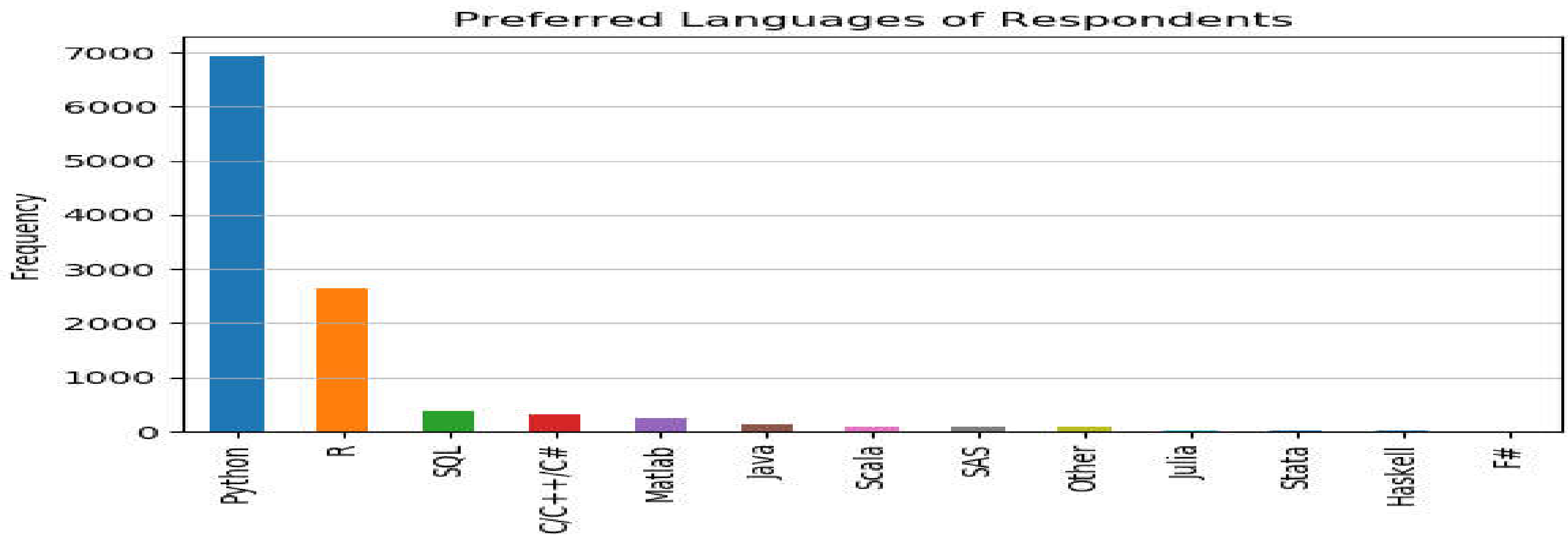


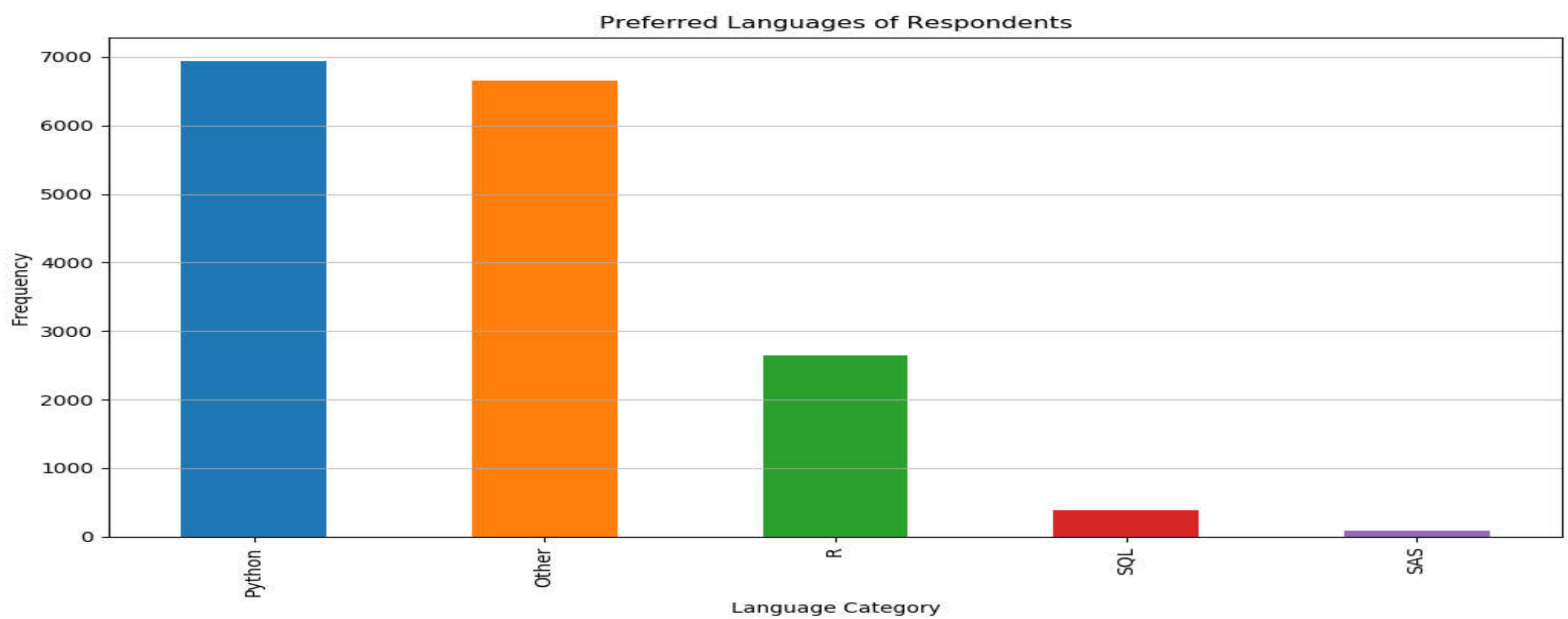
- Stage 2: plotting the variables



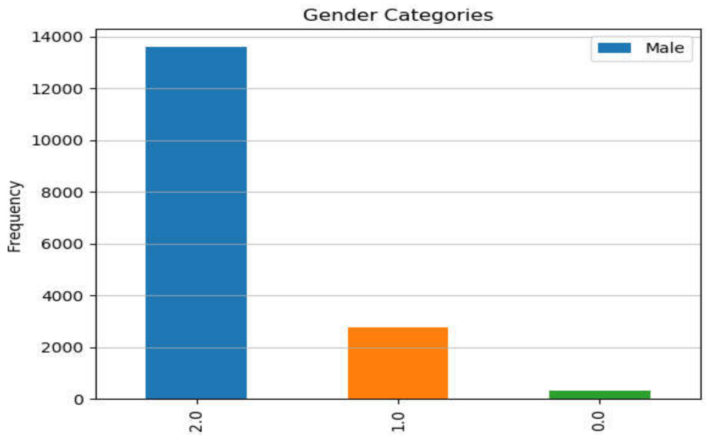


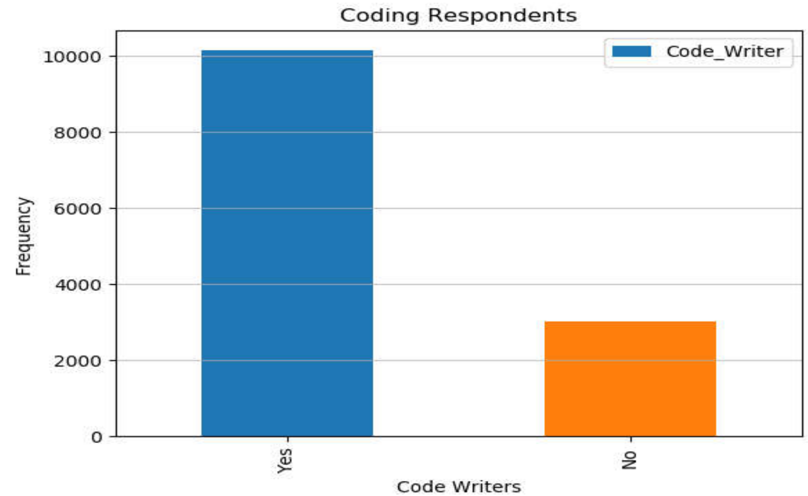
- Stage 2: plotting the variables

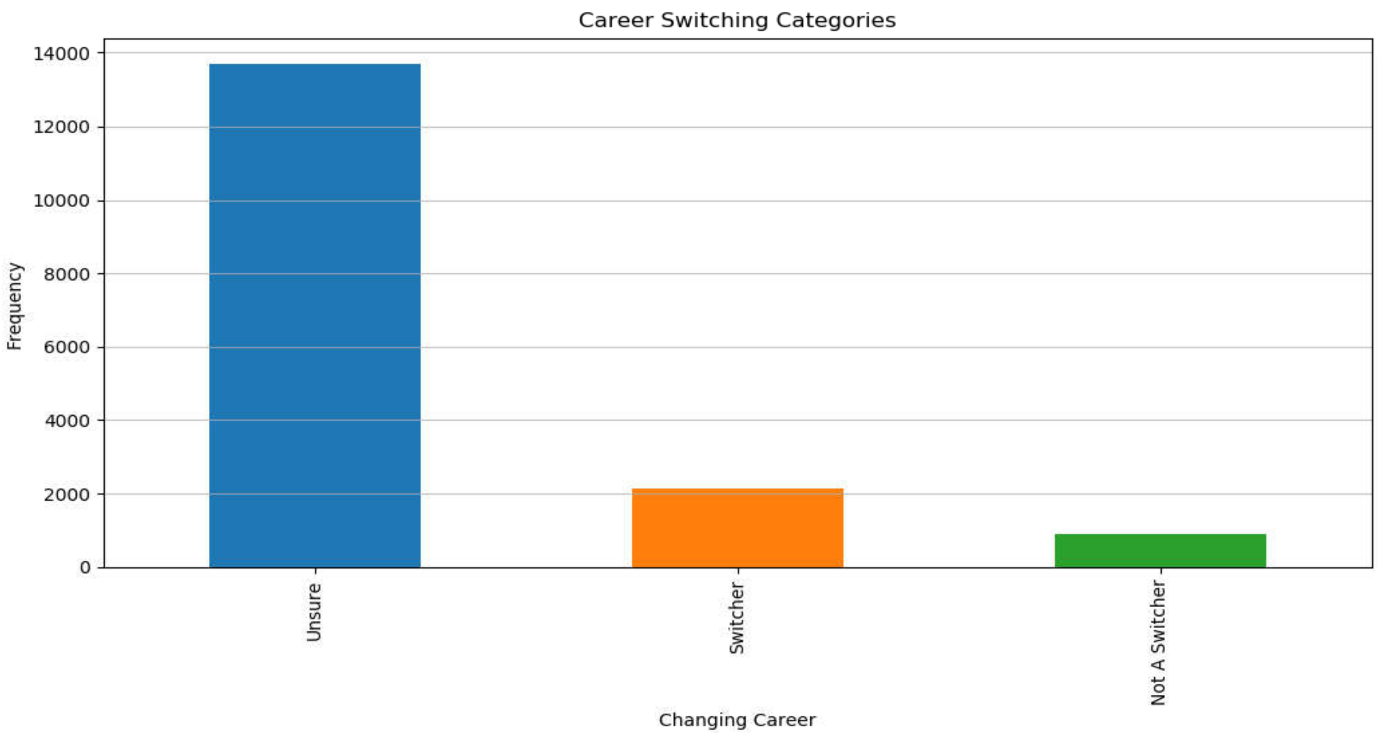


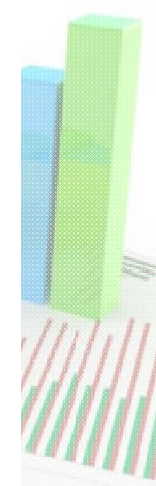


- Stage 2: plotting the variables



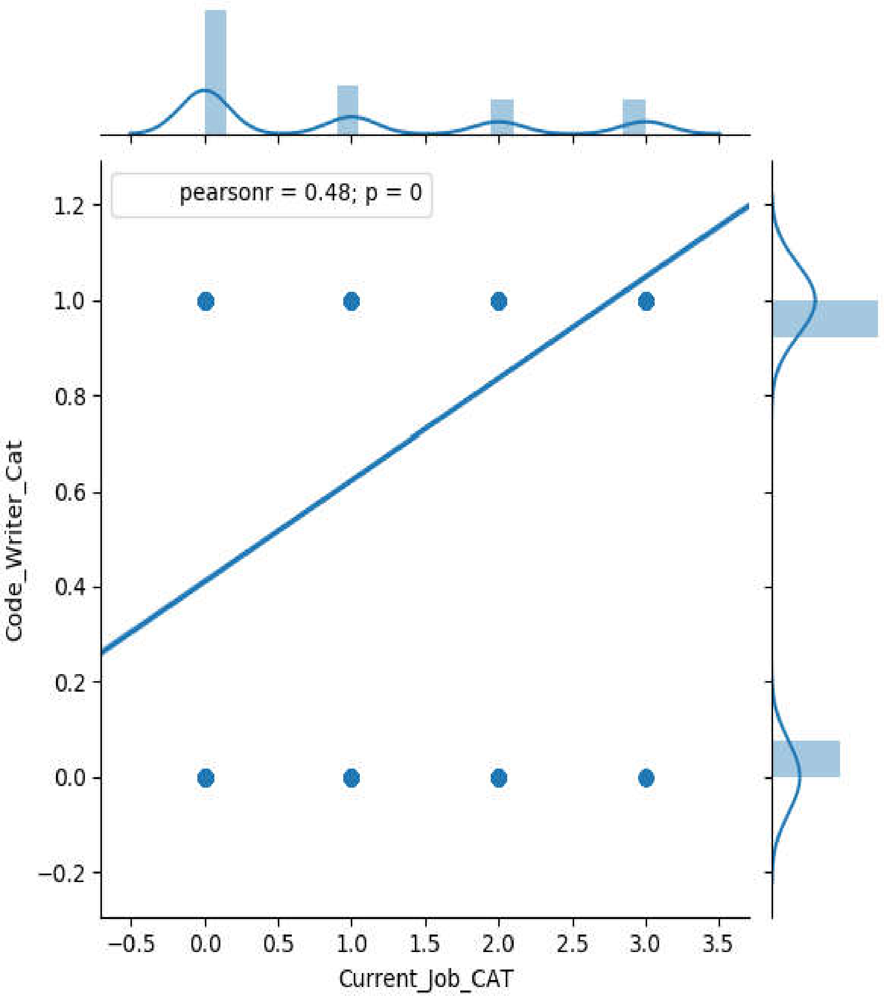


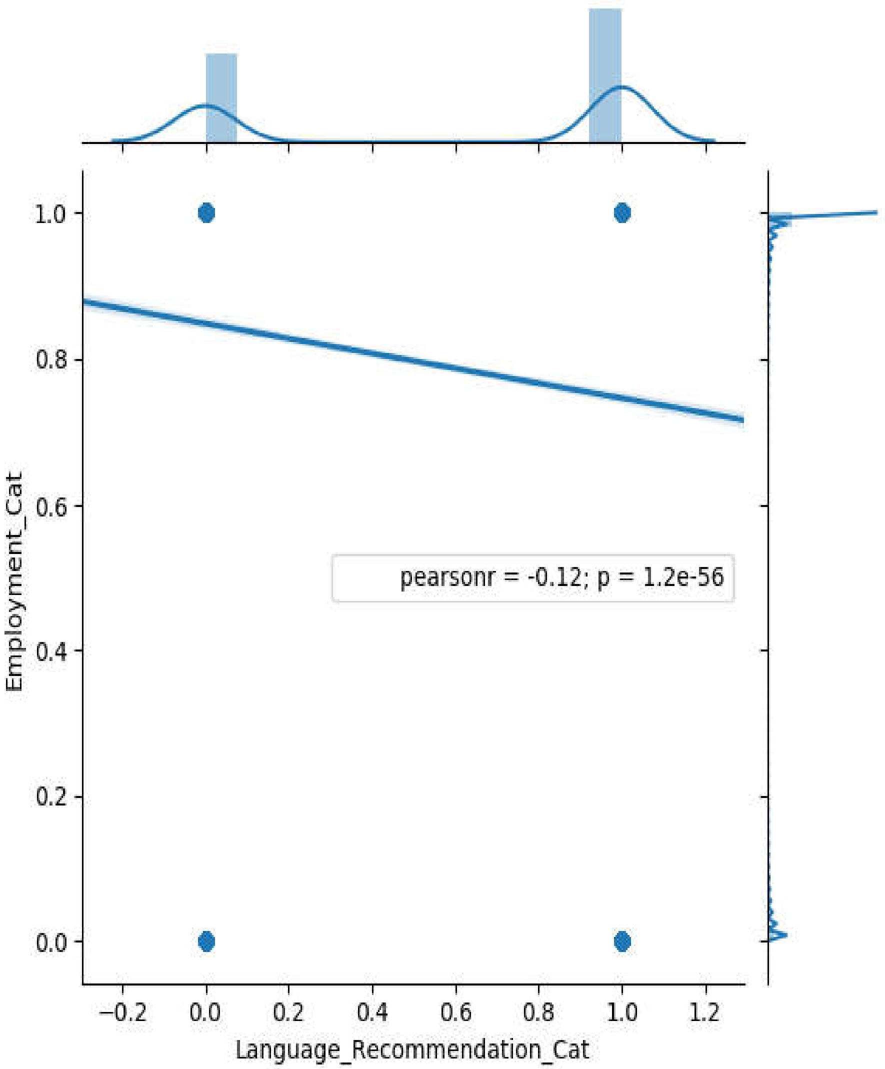




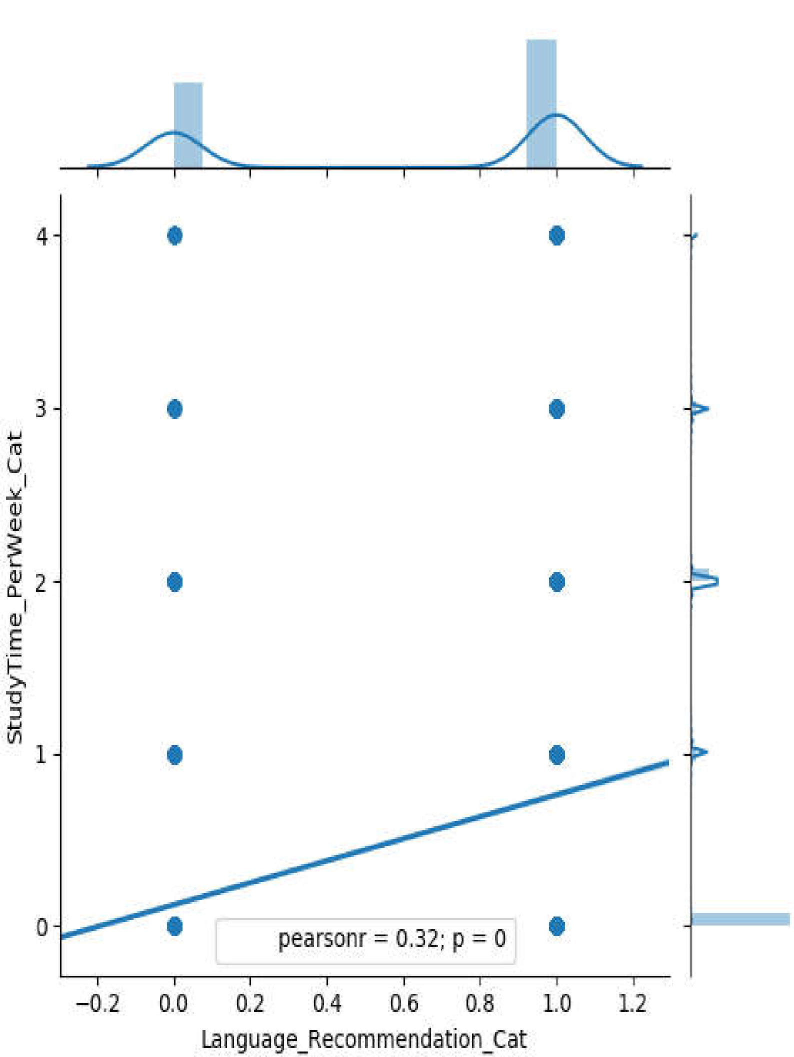
Python Methodology

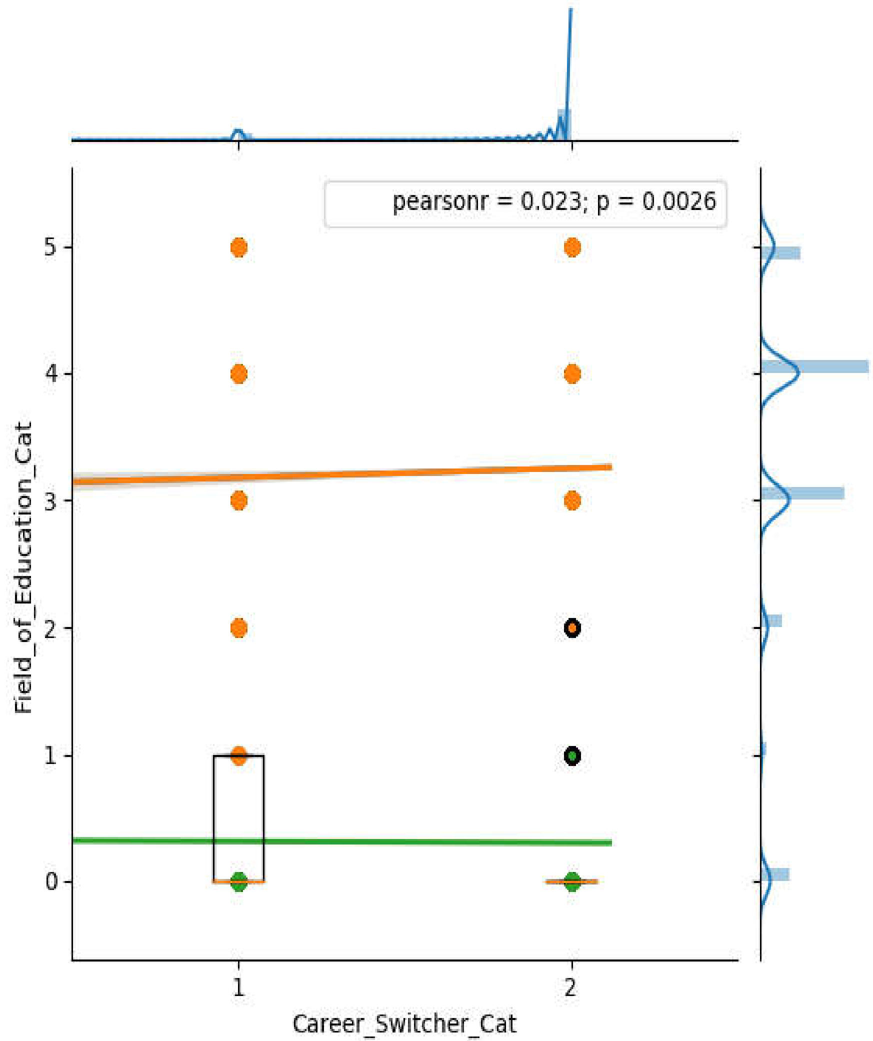
- **Stage 3: Running Statistical Tests, Scatters, and Boxplots**



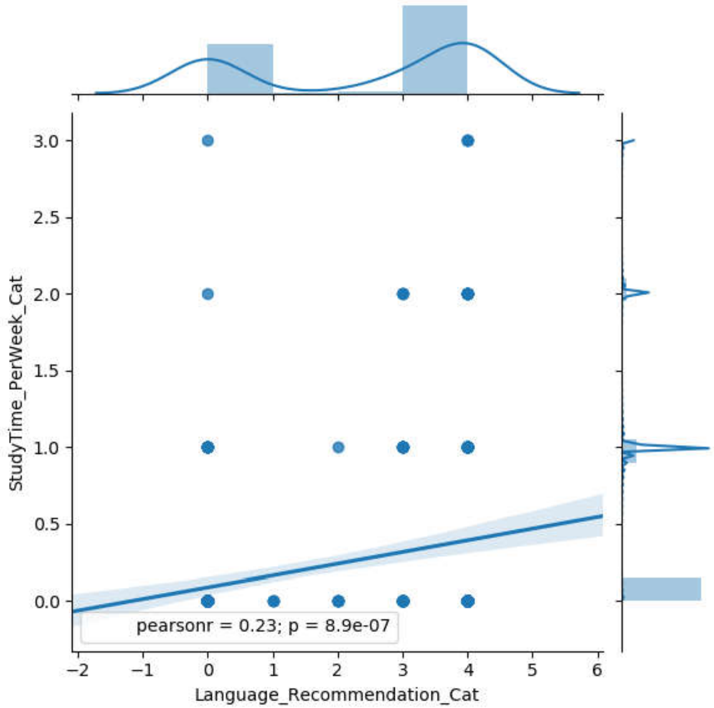


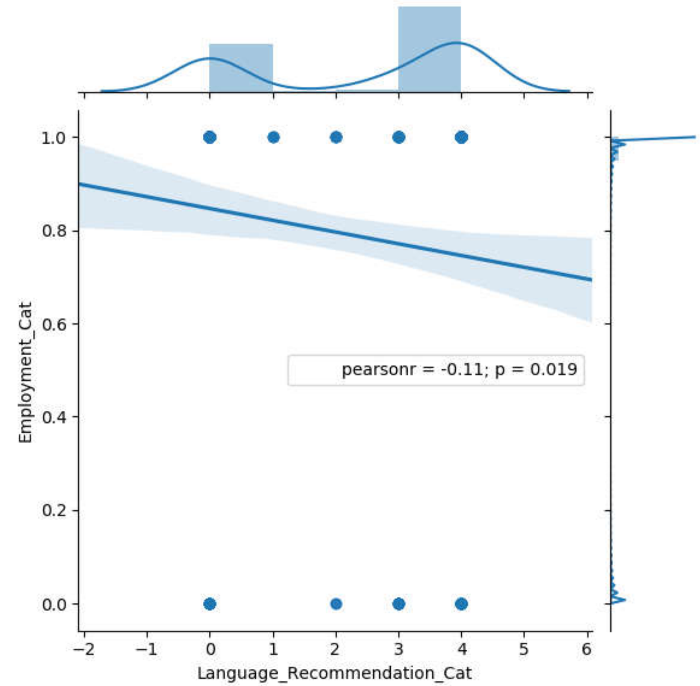
- Stage 3: Running Statistical Tests, Scatters, and Boxplots

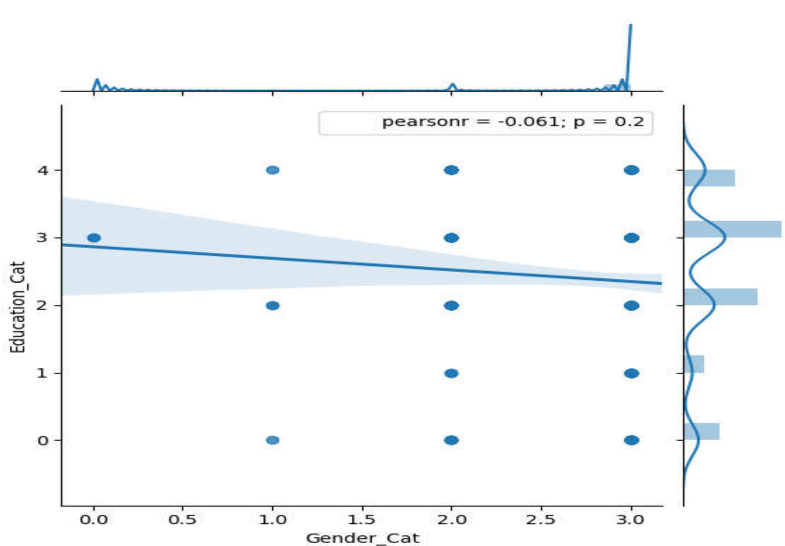


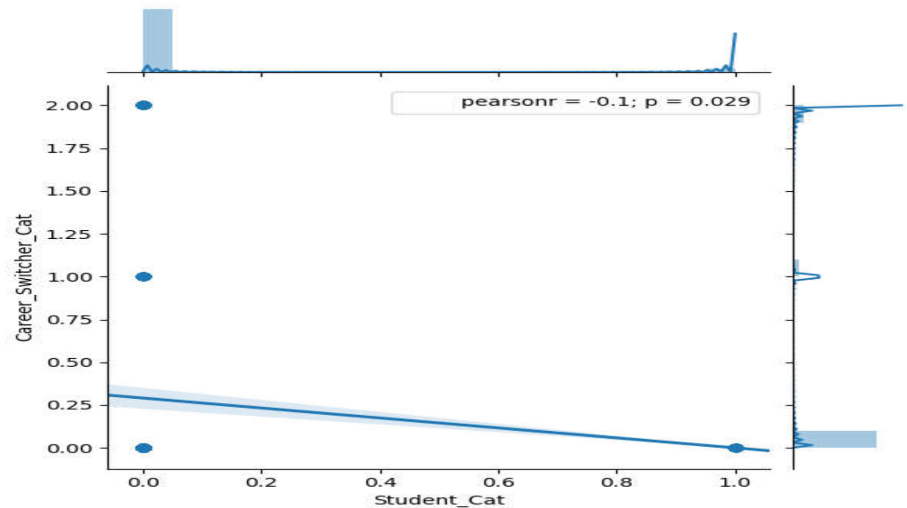


Canada Results





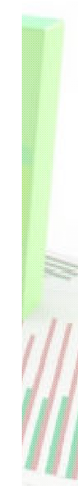




Python Regression Models

OLS Model: International

```
Dep. Variable:          Employment_Cat   R-squared:                       0.953
Model:                             OLS   Adj. R-squared:                  0.953
Method:                  Least Squares   F-statistic:                 1.318e+04
Date:                 Sun, 23 Sep 2018   Prob (F-statistic):               0.00
Time:                         11:27:57   Log-Likelihood:                 16394.
No. Observations:                16183   AIC:                        -3.274e+04
Df Residuals:                    16157   BIC:                        -3.254e+04
Df Model:                           25
Covariance Type:             nonrobust
=========================================================================================================================
                                                         coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------------------
Intercept                                              0.8874      0.010     88.858      0.000       0.868       0.907
Field_of_Education_Labeled[T.Social Science]          -0.0028      0.004     -0.771      0.441      -0.010       0.004
Field_of_Education_Labeled[T.Natural Science]          0.0006      0.002      0.255      0.799      -0.004       0.005
Field_of_Education_Labeled[T.Engineering]          -6.975e-05      0.003     -0.027      0.978      -0.005       0.005
Field_of_Education_Labeled[T.Computer Science/IT]      0.0023      0.002      1.027      0.305      -0.002       0.007
Education_Labeled[T.Some College]                      0.0641      0.004     16.439      0.000       0.056       0.072
Education_Labeled[T.Bachelor]                          0.0614      0.003     17.556      0.000       0.055       0.068
Education_Labeled[T.Master]                            0.0602      0.004     17.137      0.000       0.053       0.067
Education_Labeled[T.PH.D]                              0.0605      0.004     15.440      0.000       0.053       0.068
Gender_Labeled[T.A different identity]                 0.0284      0.012      2.327      0.020       0.004       0.052
Gender_Labeled[T.Female]                               0.0040      0.010      0.419      0.675      -0.015       0.023
Gender_Labeled[T.Male]                                 0.0038      0.010      0.401      0.688      -0.015       0.022
Age_group[T.25-34]                                    -0.0089      0.002     -4.731      0.000      -0.013      -0.005
Age_group[T.35-44]                                    -0.0090      0.002     -3.925      0.000      -0.014      -0.005
Age_group[T.45-65]                                     0.0161      0.003      6.019      0.000       0.011       0.021
Student_Labeled[T.Student]                             0.0345      0.003     10.153      0.000       0.028       0.041
StudyTime_PerWeek_Labeled[T.2 - 10 hours]              0.0265      0.003      9.715      0.000       0.021       0.032
StudyTime_PerWeek_Labeled[T.11 - 39 hours]             0.0330      0.004      8.683      0.000       0.026       0.040
StudyTime_PerWeek_Labeled[T.40+]                       0.0409      0.007      5.721      0.000       0.027       0.055
Language_Recommendation_Labeled[T.SAS]                 0.0163      0.010      1.688      0.091      -0.003       0.035
Language_Recommendation_Labeled[T.SQL]                 0.0074      0.005      1.564      0.118      -0.002       0.017
Language_Recommendation_Labeled[T.R]                   0.0043      0.002      1.954      0.051   -1.27e-05       0.009
Language_Recommendation_Labeled[T.Python]              0.0046      0.002      2.648      0.008       0.001       0.008
Code_Writer_Labeled[T.Code Writer]                    -0.9448      0.003   -357.098      0.000      -0.950      -0.940
Career_Switcher_Labeled[T.Not A Switcher]             -0.8797      0.005   -192.766      0.000      -0.889      -0.871
Career_Switcher_Labeled[T.Switcher]                   -0.9602      0.003   -353.065      0.000      -0.966      -0.955
==============================================================================
Omnibus:                      7400.296   Durbin-Watson:                   2.033
Prob(Omnibus):                   0.000   Jarque-Bera (JB):          6963311.305
Skew:                            0.687   Prob(JB):                         0.00
```



Python Regression Models

OLS Model: Canada

Model:	OLS	Adj. R-squared:	0.909
Method:	Least Squares	F-statistic:	170.7
Date:	Sun, 23 Sep 2018	Prob (F-statistic):	1.87e-195
Time:	11:32:36	Log-Likelihood:	302.89
No. Observations:	424	AIC:	-553.8
Df Residuals:	398	BIC:	-448.5
Df Model:	25		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.1776	0.130	1.364	0.173	-0.078	0.434
Field_of_Education_Labeled[T.Social Science]	0.0086	0.028	0.311	0.756	-0.046	0.063
Field_of_Education_Labeled[T.Natural Science]	-0.0180	0.020	-0.897	0.370	-0.057	0.021
Field_of_Education_Labeled[T.Engineering]	0.0008	0.022	0.036	0.971	-0.042	0.044
Field_of_Education_Labeled[T.Computer Science/IT]	0.0055	0.020	0.275	0.783	-0.034	0.045
Education_Labeled[T.Some College]	-0.1128	0.036	-3.105	0.002	-0.184	-0.041
Education_Labeled[T.Bachelor]	-0.1074	0.033	-3.270	0.001	-0.172	-0.043
Education_Labeled[T.Master]	-0.1299	0.033	-3.968	0.000	-0.194	-0.066
Education_Labeled[T.PH.D]	-0.1193	0.035	-3.388	0.001	-0.189	-0.050
Gender_Labeled[T.A different identity]	-0.0281	0.141	-0.199	0.842	-0.305	0.249
Gender_Labeled[T.Female]	0.0188	0.126	0.148	0.882	-0.230	0.267
Gender_Labeled[T.Male]	-0.0032	0.126	-0.026	0.980	-0.250	0.244
Age_group[T.25-34]	0.0291	0.018	1.586	0.114	-0.007	0.065
Age_group[T.35-44]	0.0322	0.019	1.665	0.097	-0.006	0.070
Age_group[T.45-65]	-0.0239	0.021	-1.155	0.249	-0.065	0.017
Student_Labeled[T.Student]	-0.0289	0.031	-0.927	0.355	-0.090	0.032
StudyTime_PerWeek_Labeled[T.2 - 10 hours]	-0.0454	0.026	-1.773	0.077	-0.096	0.005
StudyTime_PerWeek_Labeled[T.11 - 39 hours]	-0.0479	0.035	-1.353	0.177	-0.118	0.022
StudyTime_PerWeek_Labeled[T.40+]	-0.0461	0.061	-0.753	0.452	-0.166	0.074
Language_Recommendation_Labeled[T.SAS]	0.0285	0.064	0.448	0.654	-0.096	0.153
Language_Recommendation_Labeled[T.SQL]	-0.1090	0.046	-2.378	0.018	-0.199	-0.019
Language_Recommendation_Labeled[T.R]	0.0032	0.020	0.157	0.876	-0.037	0.043
Language_Recommendation_Labeled[T.Python]	-0.0097	0.015	-0.629	0.530	-0.040	0.021
Code_Writer_Labeled[T.Code Writer]	0.9211	0.025	36.803	0.000	0.872	0.970
Career_Switcher_Labeled[T.Not A Switcher]	0.7834	0.039	20.284	0.000	0.707	0.859
Career_Switcher_Labeled[T.Switcher]	0.9531	0.025	37.873	0.000	0.904	1.003

Omnibus:	296.334	Durbin-Watson:	1.994
Prob(Omnibus):	0.000	Jarque-Bera (JB):	32137.207
Skew:	-2.085	Prob(JB):	0.00
Kurtosis:	45.446	Cond. No.	75.9

Python Regression Models

GLM Model: Canada

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:        Employment_Cat   No. Observations:                  424
Model:                           GLM   Df Residuals:                      398
Model Family:               Binomial   Df Model:                           25
Link Function:                 logit   Scale:                          1.0000
Method:                         IRLS   Log-Likelihood:                    nan
Date:               Sun, 23 Sep 2018   Deviance:                          nan
Time:                       11:43:06   Pearson chi2:                     16.0
No. Iterations:                  100   Covariance Type:             nonrobust
==================================================================================================================
                                                     coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------------
Intercept                                       -166.4406   6.95e+07   -2.4e-06      1.000   -1.36e+08    1.36e+08
Field_of_Education_Labeled[T.Social Science]       0.4852   1.42e+07   3.41e-08      1.000   -2.79e+07    2.79e+07
Field_of_Education_Labeled[T.Natural Science]    -66.4493   9.55e+06  -6.96e-06      1.000   -1.87e+07    1.87e+07
Field_of_Education_Labeled[T.Engineering]          1.5665   1.14e+07   1.37e-07      1.000   -2.24e+07    2.24e+07
Field_of_Education_Labeled[T.Computer Science/IT]  1.5764   1.06e+07   1.49e-07      1.000   -2.07e+07    2.07e+07
Education_Labeled[T.Some College]                -98.1794   1.73e+07  -5.69e-06      1.000   -3.38e+07    3.38e+07
Education_Labeled[T.Bachelor]                   -101.6872   1.51e+07  -6.73e-06      1.000   -2.96e+07    2.96e+07
Education_Labeled[T.Master]                     -166.4213   1.49e+07  -1.12e-05      1.000   -2.92e+07    2.92e+07
Education_Labeled[T.PH.D]                       -132.2648   1.57e+07  -8.41e-06      1.000   -3.08e+07    3.08e+07
Gender_Labeled[T.A different identity]            33.0635   7.69e+07    4.3e-07      1.000   -1.51e+08    1.51e+08
Gender_Labeled[T.Female]                         133.2902    6.9e+07   1.93e-06      1.000   -1.35e+08    1.35e+08
Gender_Labeled[T.Male]                           131.4642   6.86e+07   1.92e-06      1.000   -1.34e+08    1.34e+08
Age_group[T.25-34]                                68.1588   8.31e+06    8.2e-06      1.000   -1.63e+07    1.63e+07
Age_group[T.35-44]                                69.1317   9.32e+06   7.42e-06      1.000   -1.83e+07    1.83e+07
Age_group[T.45-65]                               -65.3287   8.28e+06  -7.89e-06      1.000   -1.62e+07    1.62e+07
Student_Labeled[T.Student]                        -3.5095   1.66e+07  -2.12e-07      1.000   -3.24e+07    3.24e+07
StudyTime_PerWeek_Labeled[T.2 - 10 hours]         -0.8517    1.3e+07  -6.57e-08      1.000   -2.54e+07    2.54e+07
StudyTime_PerWeek_Labeled[T.11 - 39 hours]        29.2432   1.86e+07   1.57e-06      1.000   -3.65e+07    3.65e+07
StudyTime_PerWeek_Labeled[T.40+]                  91.0618   3.32e+07   2.74e-06      1.000   -6.51e+07    6.51e+07
Language_Recommendation_Labeled[T.SAS]             2.4983   3.46e+07   7.22e-08      1.000   -6.79e+07    6.79e+07
Language_Recommendation_Labeled[T.SQL]           -33.2025   1.38e+07  -2.41e-06      1.000    -2.7e+07     2.7e+07
Language_Recommendation_Labeled[T.R]              33.5419   1.04e+07   3.23e-06      1.000   -2.04e+07    2.04e+07
Language_Recommendation_Labeled[T.Python]        -32.4392   7.85e+06  -4.13e-06      1.000   -1.54e+07    1.54e+07
Code_Writer_Labeled[T.Code Writer]               333.2502   1.18e+07   2.82e-05      1.000   -2.31e+07    2.31e+07
Career_Switcher_Labeled[T.Not A Switcher]        102.3846   1.08e+07   9.49e-06      1.000   -2.11e+07    2.11e+07
Career_Switcher_Labeled[T.Switcher]              400.9001   1.29e+07   3.12e-05      1.000   -2.52e+07    2.52e+07
==================================================================================================================
```

Using Models to Predict

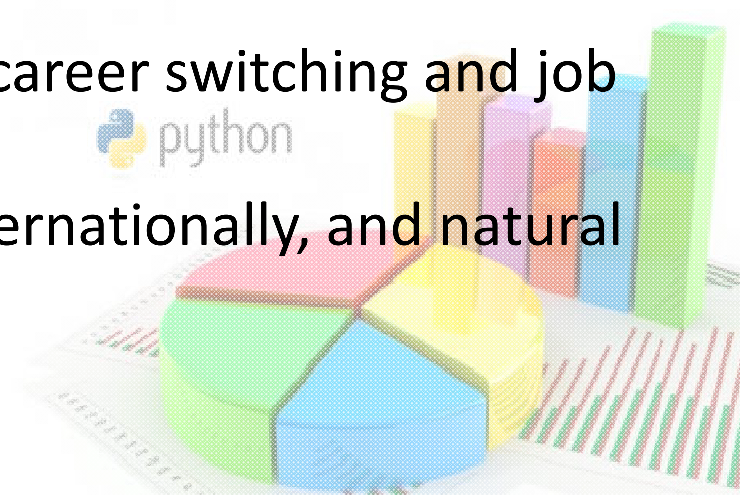
```
In [684]: model =
sm.OLS(Data_Scientist['Employment_Cat'],
Data_Scientist['Gender_Cat']).fit()
     ...: predictions =
model.predict(Data_Scientist['Gender_Cat'])
     ...: model.summary()
     ...: print(predictions.head(10))
0    0.000000
1    0.145468
2    0.218203
3    0.218203
4    0.218203
5    0.218203
6    0.218203
7    0.145468
8    0.145468
9    0.218203
dtype: float64
```

```
In [686]: model = sm.OLS(Data_Scientist['Employment_Cat'],
Data_Scientist['Language_Recommendation_Cat']).fit()
     ...: predictions =
model.predict(Data_Scientist['Language_Recommendation_Cat'])
     ...: model.summary()
     ...: print(predictions.head(10))
0    0.000000
1    0.276806
2    0.207605
3    0.276806
4    0.276806
5    0.276806
6    0.207605
7    0.138403
8    0.276806
9    0.276806
dtype: float64
```

```
In [687]: model = sm.OLS(Data_Scientist['Employment_Cat'],
Data_Scientist['Education_Cat']).fit()
     ...: predictions =
model.predict(Data_Scientist['Education_Cat'])
     ...: model.summary()
     ...: print(predictions.head(10))
0    0.129915
1    0.194873
2    0.194873
3    0.194873
4    0.259831
5    0.259831
6    0.194873
7    0.129915
8    0.129915
9    0.129915
dtype: float64
```

Key Findings

- Surprisingly, Canada's market for Data Science and Analysis does not require formal education, but perhaps more professional designations. This is unlike the international market that requires formal education with tendency to require more Master's degrees.
- There is a slight increase in tendency for hiring female D.S (although insignificant).
- Surprisingly, Canadian market appeals more to R and SAS than the International market appealing to python.
- The international market most fitted age groups is (45-65), while Canada' s age group is (25-44).
- There is a positive association between career switching and job employment.
- Social Science is the least to be hired internationally, and natural science is the least in Canada.




WHEN
WHO
WHERE
WHAT
WHY
HOW
?