Data Analysis Capstone Project

Helena Rolle,

Data Analyst/Developer

helenarolle@bah.gov.bs

June 14th, 2022

IBM Developer

SKILLS NETWORK

Table of Contents

- EXECUTIVE SUMMARY
- INTRODUCTION
- METHODOLOGY
- RESULTS: Data Collection
- RESULTS: Data Organization
- RESULTS: Webscraping US IT Market
- RESULTS: Data Analysis
- RESULTS: Visual Analysis
- RESULTS: Computational Analysis
- RESULTS: Relational Analysis
- RESULTS: Trend Analysis – Line Chart
- PROGRAMMING LANGUAGE TRENDS: top 5 Language
- DATABASE TRENDS: top 5 Databases

DATABASE TRENDS - FINDINGS & IMPLICATIONs
TRENDS IN PLATFORMS – visual comparison
TRENDS IN WEBFRAME- visual comparison
DASHBOARD Link
DASHBOARD 1: Current Technology Usage
DASHBOARD 2: Future Technology Usage
DASHBOARD 3: Demographics
COURSE LINKS - GitHub
DISCUSSION: Analytical Findings
DISCUSSION: Computation & Trend Analysis
DISCUSSION: Prog.Language Analysis
DISCUSSION: Database Analysis
DISCUSSION: WebFrame Trends
OVERALL FINDINGS & IMPLICATIONS
CONCLUSION: Findings
CONCLUSION: Future Analyses
APPENDIX: Correlation – Regression Line
APPENDIX – Impact of Multiple Variables

Table of Contents

- APPENDIX: Correlation – Regression Line
- APPENDIX: Categorizing IT Tasks
- APPENDIX: All Current Language Trends
- APPENDIX: All Database Trends
- APPENDIX – Video: How to Get Technology Job
- APPENDIX – Links to IT Datasets
- APPENDIX – Data Dictionary 1/6
- ACKNOWLEDGEMENTS
- REFERENCES
- ANIMATED THANK YOU
- PRESENTATION APPLAUSE
- Close Slide

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

IBM Developer
SKILLS NETWORK

EXECUTIVE SUMMARY

- Data Constituent

1. Most data variables captured as string objects; 4 typed as float64
2. Higher number job openings in US northern cities eg capital (5,316); lowest openings in Austin (434)
3. Ratio of male to female respondents: 10,480 (93.5%):731(6.5%) but median salary $57,745 for both groups; genderqueers has median salary $66,000
4. More middle aged millenniums involved in IT – median age range 29 years and 75th percentile at 34 years
5. Over 900 salaries abnormalities found beyond Q3 point, $193,976
6. Most respondents are true developers compared to ones code periodically in ratio approx. 10:1
7. All techs work more hours in comparison to doing code reviews; younger ones spend longer hours at code reviews

EXECUTIVE SUMMARY

- Languages
  1. Well paid opportunities for professionals: salary range from $84,727 (PHP) to $130,801 (Swift)
  2. C language in highest demand in US(13,498) followed in 2nd distance by Java (2609) and MySQLServer with 0
  3. Scrapped Data: Python is only language with both high openings (1173) and high average salary ($114,383)
  4. C++ has low job opening but high average annual salary

- Databases
  1. Slight increase use of PostgreSql and MongoDB for the future
  2. Current leader, MySql and MS SqlServer will have questionable fate; MySql significant decrease while MS SqlServer is not forecast as a future choice
  3. New databases – Redis & Elastsearch are cited as moderate future starts but currently not in use
  4. Postgresql, Sqlite & MongoDB showed slight increase preference for future use

EXECUTIVE SUMMARY

- Platforms

1. Seeming declining preference for Windows as Linux and Docker cited to have higher usage
2. Introduction of Slack (a messaging app for business that connects people to information) as future technology
3. Arduino and Heroku seen as competing forces

- WebFrames

1. Current leaders:jQuery, React.js, ASP.net & Angular as noted by bubble size
2. From eyeball view, React.js, ASP.net & Angular are competitors and voted to retain constant future use
3. React.js now poised to become leader with decline of jQuery
4. Vue.js webframe to show stronger presence (increase use)

INTRODUCTION

- Nature of Analysis
  - Collection of data
  - Cleaning of data
  - Exploration & Visualization of data for insights
  - Interpretation of data for inferences
  - Presentation of the data analysis
- Problem Statement: To analyze current and future , landscapes of IT
- Audience: Any IT enthusiast, Trend follower or Job Seeker
- Arising Questions
  1. How are data/variables sourced, typed & distributed?
  2. What are some US based IT opportunities?
  3. How are IT personnel remunerated?
  4. Are there noted characteristics for IT specialists age, jobFactor, WorkLoc, etc?
  5. What are sentiments of IT personnel on future technology – Languages, Database, Platform, WebFrame?

METHODOLOGY

- Data Sources
  1. GitHub Platform
  2. Kaggle job site under Public Domain
  3. Survey data: 2019 Stackoverflow blog
- Analysis Method
  1. Data Distribution Visuals:
     - Histogram
     - Bar chart & Column chart
     - Box Plot
     - Pie chart
     - Line chart
     - WordCloud
  2. Statistical Methods
     - Summaries of central tendencies
     - Correlation

RESULTS: Data Collection

1. Data Collection Details
   1. Via API from Kaggle.com & GitHub.com (FLASK)
      a. Requests.get payload parameter using key-value pair
      b. Viewed results in json format over HTTP
      c. Results outputted in .csv format

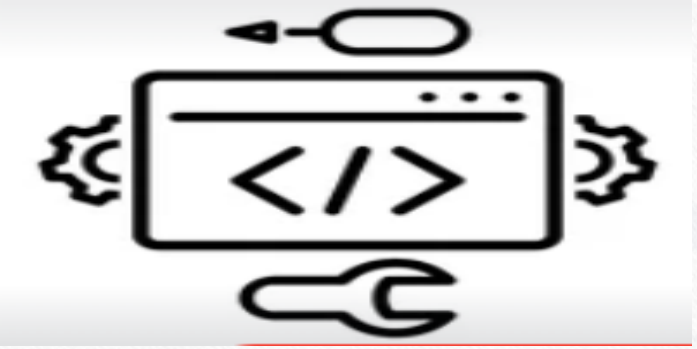
   2. Via Web Scraping:
      a. Requests.get as BeautifulSoup object

   3. Survey
      a. Normalized and unnormalized datasets in .csv format

RESULTS: Data Organization

Web-Based

1. Table format: html style (tr,td tags)
2. Scrapped from .json format to .csv file



Survey

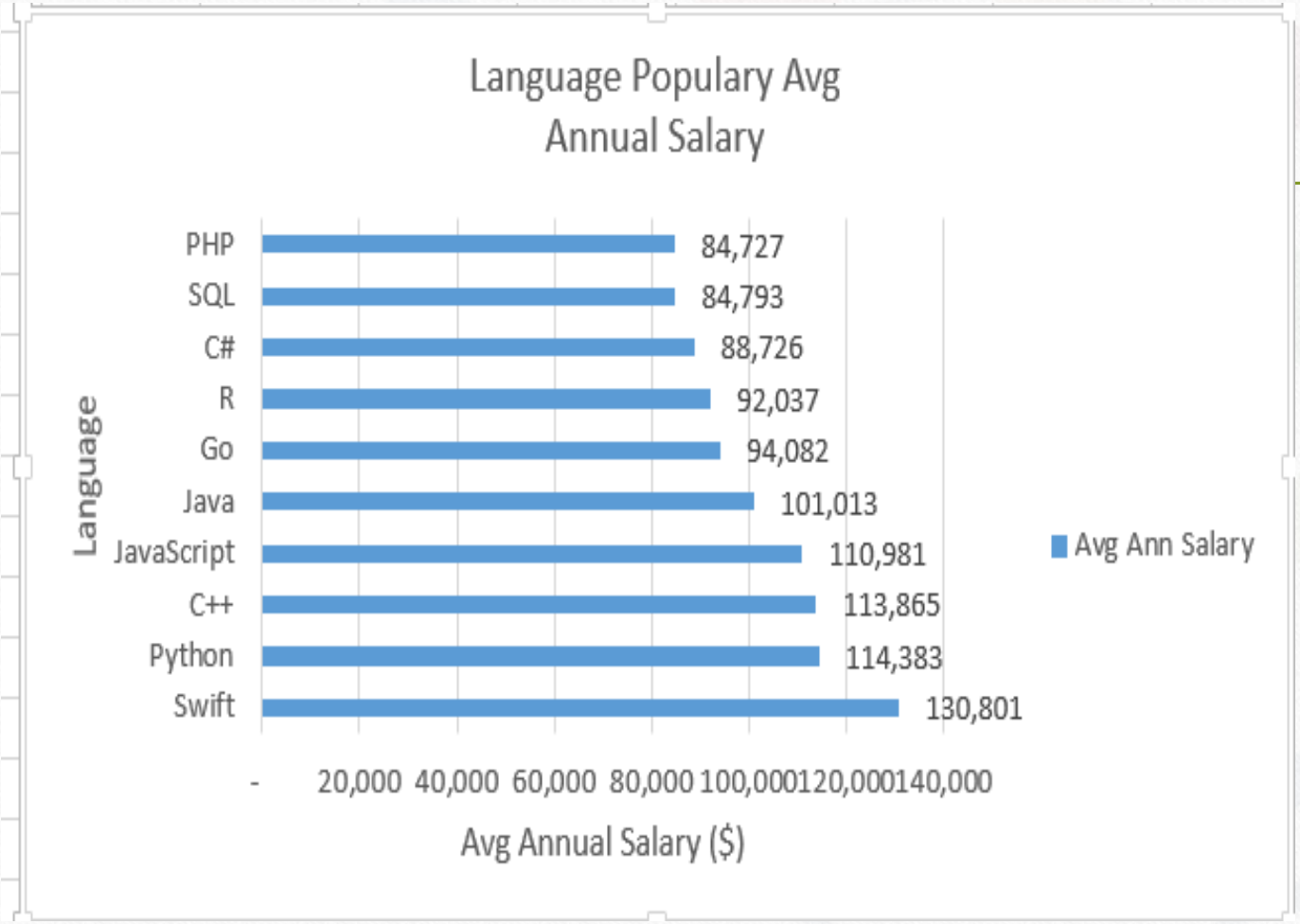
1. Table format: Rows & Columns
2. 90,000 responses from different sexual demographics
3. Dataset shape: 11552 by 85 with datatype mainly of string objects and 4 float64 columns
4. Data from 135 countries

RESULTS – Survey respondents

Number of Only Man/Male Responders is Man	10480
Woman	731
Non-binary, genderqueer, or gender non-conforming	63
Man;Non-binary, genderqueer, or gender non-conforming	26
Woman;Non-binary, genderqueer, or gender non-conforming	14
Woman;Man	9
Woman;Man;Non-binary, genderqueer, or gender non-conforming	2
Name: Gender, dtype: int64	

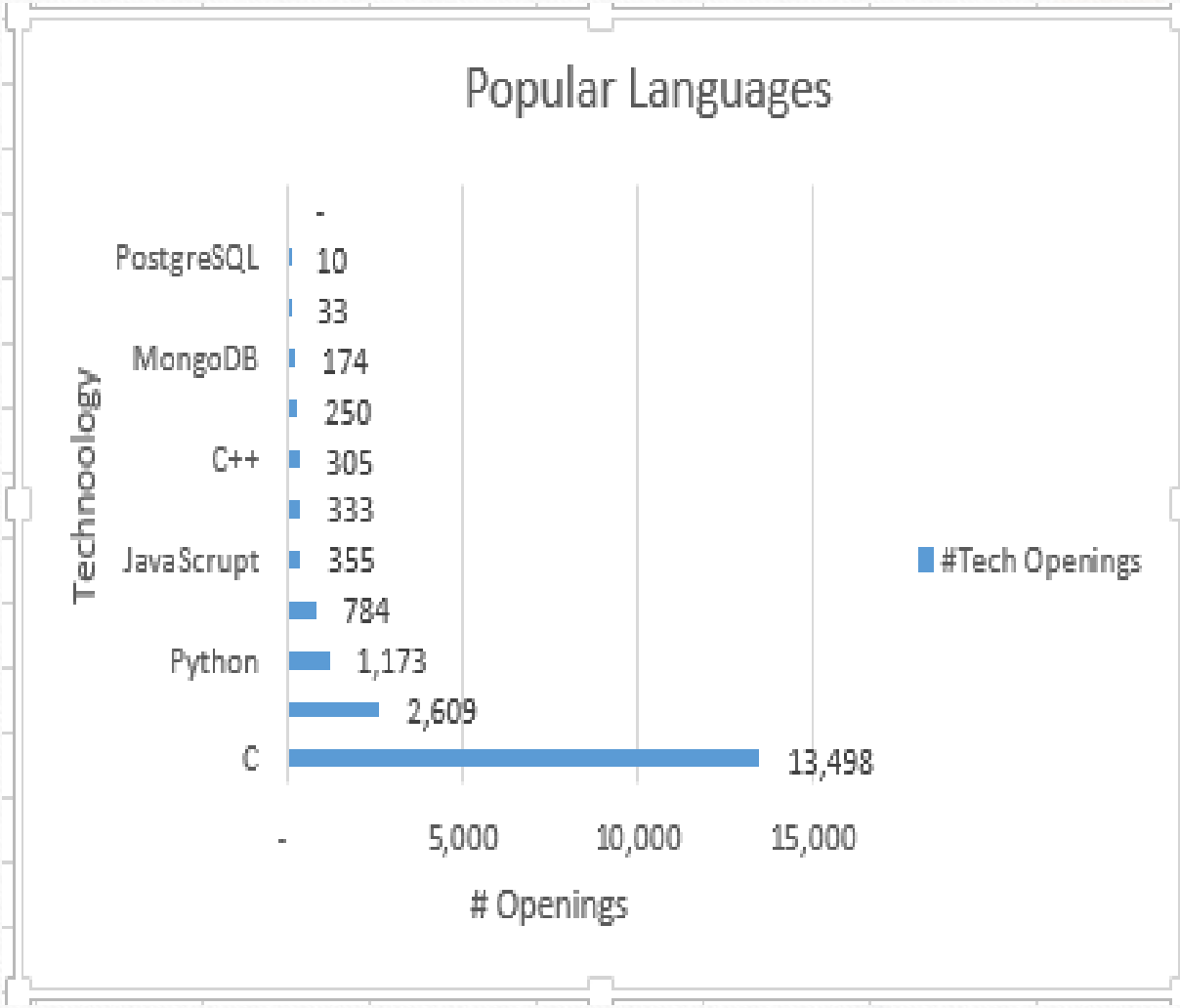
1. Number of male survey responders out number females by far – 10,480: 731
2. Though numerically low, IT field includes other gender types: genderqueers, non-binary

RESULTS: Webscraping IT Language Salary



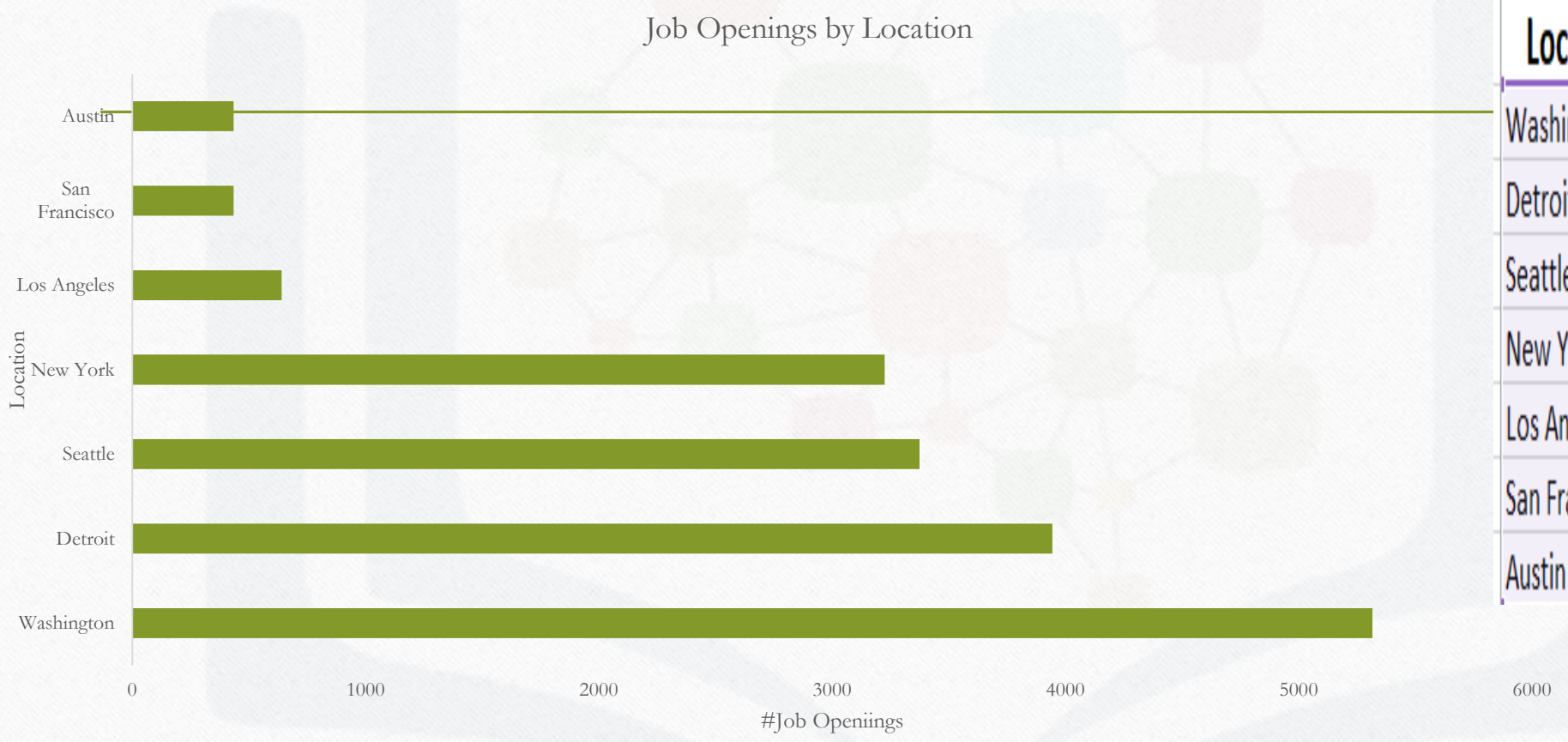
Language	Avg. Annual Salary
Swift	$130,801
Python	$114,383
C++	$113,865
Javascript	$110,981
Java	$101,013
Go	$94,082
R	$92,037
C#	$88,726
SQL	$84,793
PHP	$84,727

RESULTS: API Job Postings



Technology	#Tech Openings
C	13,498
Java	2,609
Python	1,173
Oracle	784
JavaScript	355
C#	333
C++	305
SQL Server	250
MongoDB	174
Scala	33
PostgreSQL	10
MySQL Server	-

RESULTS: API US IT Market Access



Location	#Job Openings
Washington	5,316
Detroit	3,945
Seattle	3,375
New York	3,226
Los Angeles	640
San Francisco	435
Austin	434

RESULTS: Data Analysis

a. Exploratory Analysis

1. Dataset explored to identify and remove duplicates

```
Respondent        154
MainBranch        154
Hobbyist          154
OpenSourcer       154
OpenSource        154
                 ...
Sexuality         149
Ethnicity         146
Dependents        150
SurveyLength      154
SurveyEase        154
Length: 85, dtype: int64 is Number of Duplicates
```

```
# your code goes here

print("Number of Rows before Removing Duplicates: " + str(len(df)))
print("Number of Rows after Removing Duplicates: " + str(len(Survey_ConDta)))
```

```
Number of Rows before Removing Duplicates: 11552
Number of Rows after Removing Duplicates: 11552
```

RESULTS: Data Analysis

b. Explored to identify and impute missing values based on datatype
   - Numeric: mean/median replacement
   - Object(String) mode frequency

- Before Imputation

```
# your code goes here

missWorkLoc = Survey_ConDta['WorkLoc'].isnull().sum()
print("Missing Values in WorkLoc Column is: " + str(missWorkLoc))
```

```
Missing Values in WorkLoc Column is: 32
```

- After Imputation

```
# your code goes here

Survey_ConDta['WorkLoc'].isnull().sum()
```

```
0
```

RESULTS: Data Analysis

c. Explore to Normalize data for level comparisons

```
# ANALAZING COMPTOTAL FIELD
Survey_ConDta['CompTotal']

#Number of Missing values in CompTotal column
```

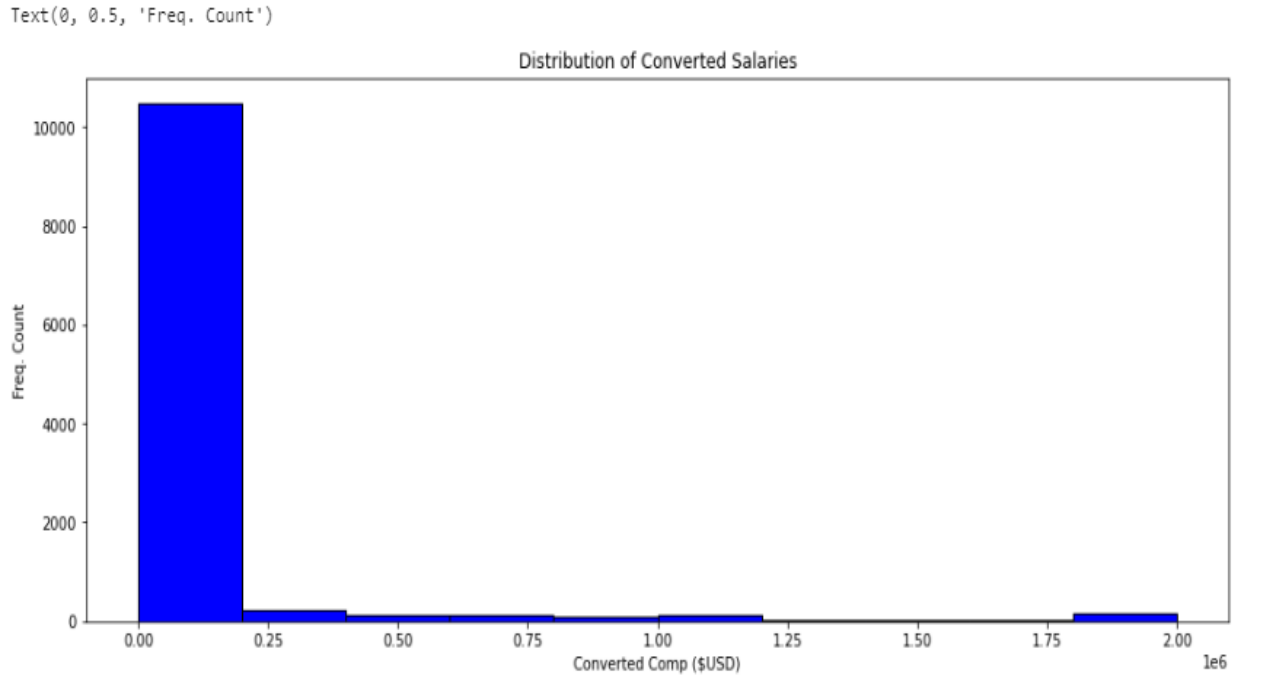
```
0         61000.0
1        138000.0
2         90000.0
3         29000.0
4         90000.0
           ...
11547    130000.0
11548     74400.0
11549    105000.0
11550     80000.0
11551         NaN
Name: CompTotal, Length: 11398, dtype: float64
```

```
Survey_ConDta['NormalAnnualCompensation']
```

```
0        61000.000000
1       138000.000000
2        90000.000000
3       348000.000000
4        90000.000000
            ...
11547   130000.000000
11548    74400.000000
11549   105000.000000
11550    80000.000000
11551   749993.196424
Name: NormalAnnualCompensation, Length: 11552, dtype: float64
```

RESULTS: Visual Analysis

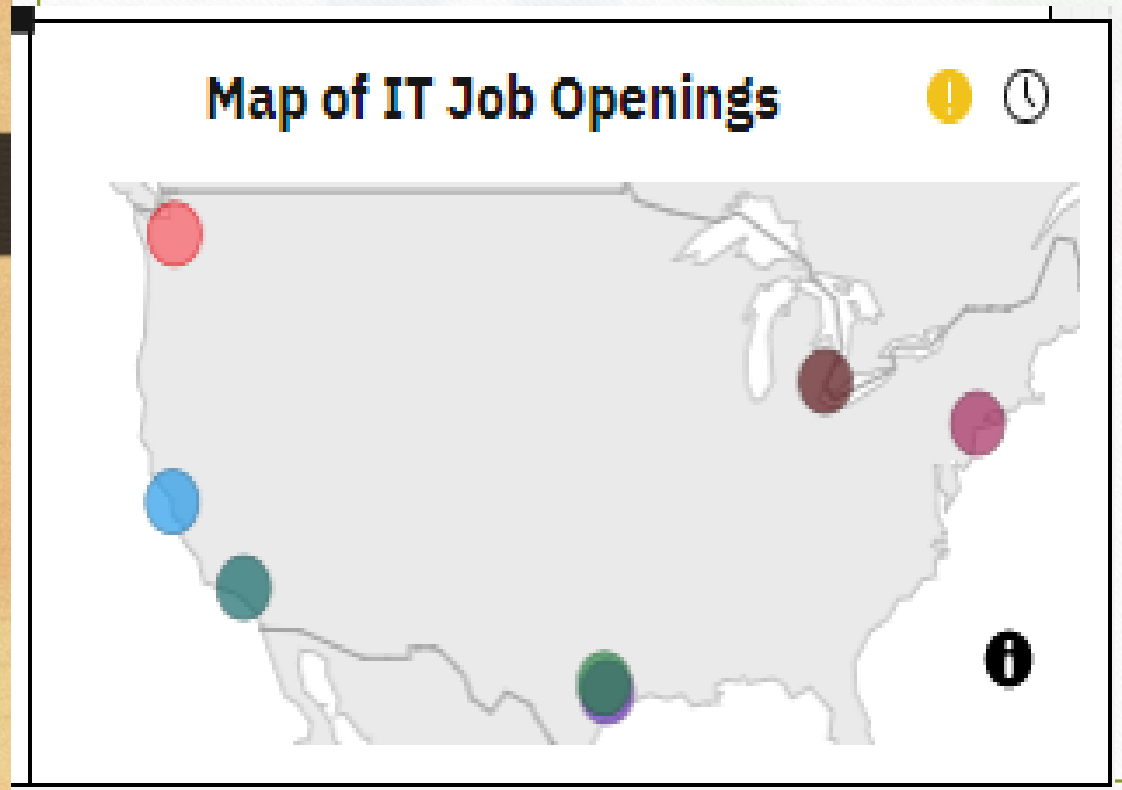
b. 1. Histogram showing Distribution: How is the *ConvertedComp* distributed?



1. Salary has been normalized so that annual salary (USD) is obtained regardless of pay frequency and amounts;

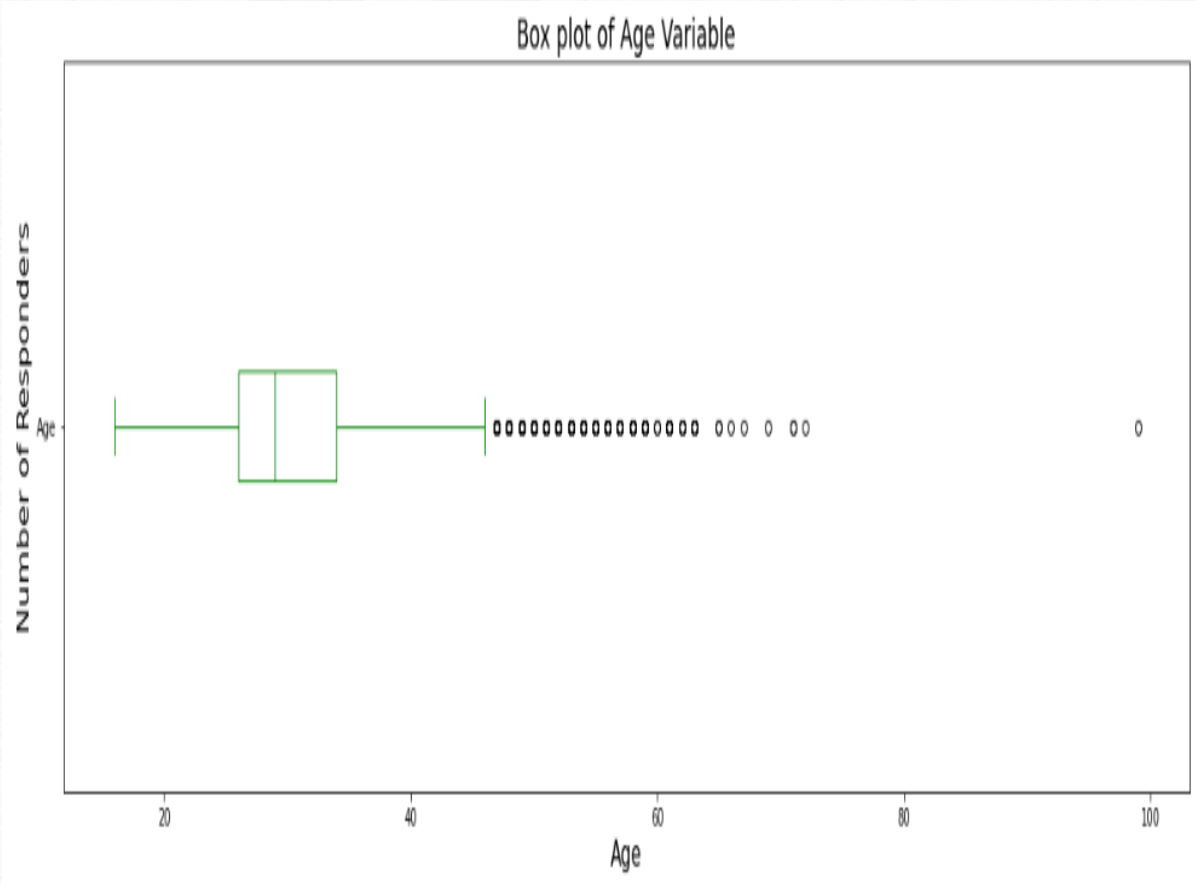
2. Right skewed distribution: Many lower salaries and fewer higher salaries

RESULTS: Visual Analysis



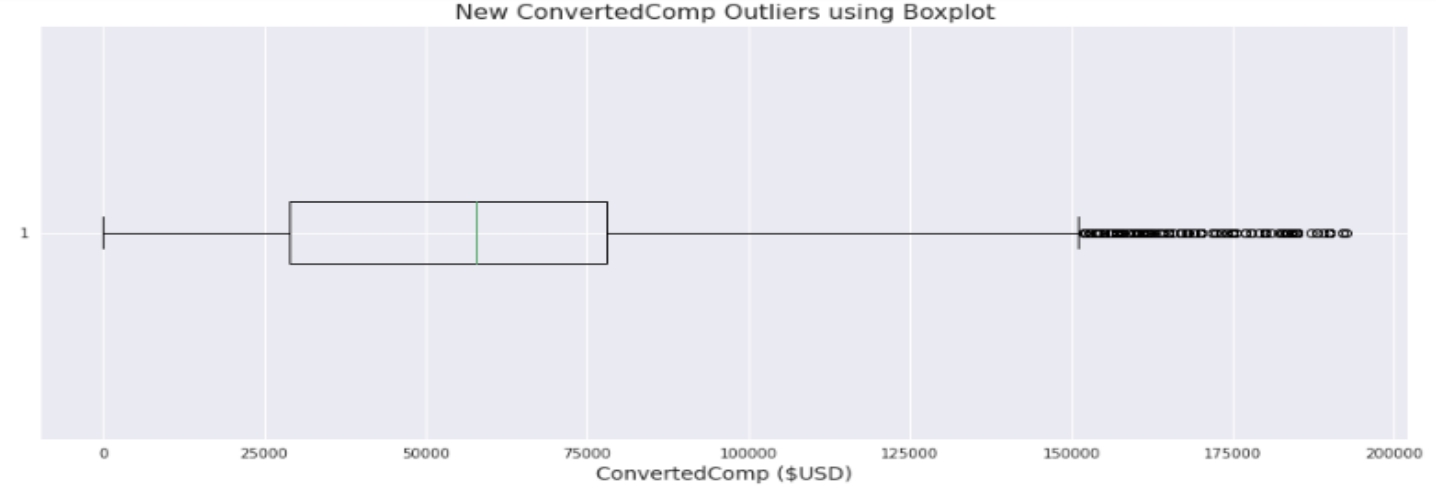
1. Map showing US Job Openings Distribution: How are jobs openings distributed in US?
2. Notably most openings are in northern US cities – Washington recorded highest of 5,316

RESULTS: Visual Analysis



1. Most IT respondents younger: median age of 29 years
2. Single extreme age of 100 years – may or may not be true outlier
3. 50% of age range between 25 – 35
4. Since the median line is closer to Q1, distribution is positively skewed: mean > median with higher frequency of higher values

RESULTS: Visual Analysis



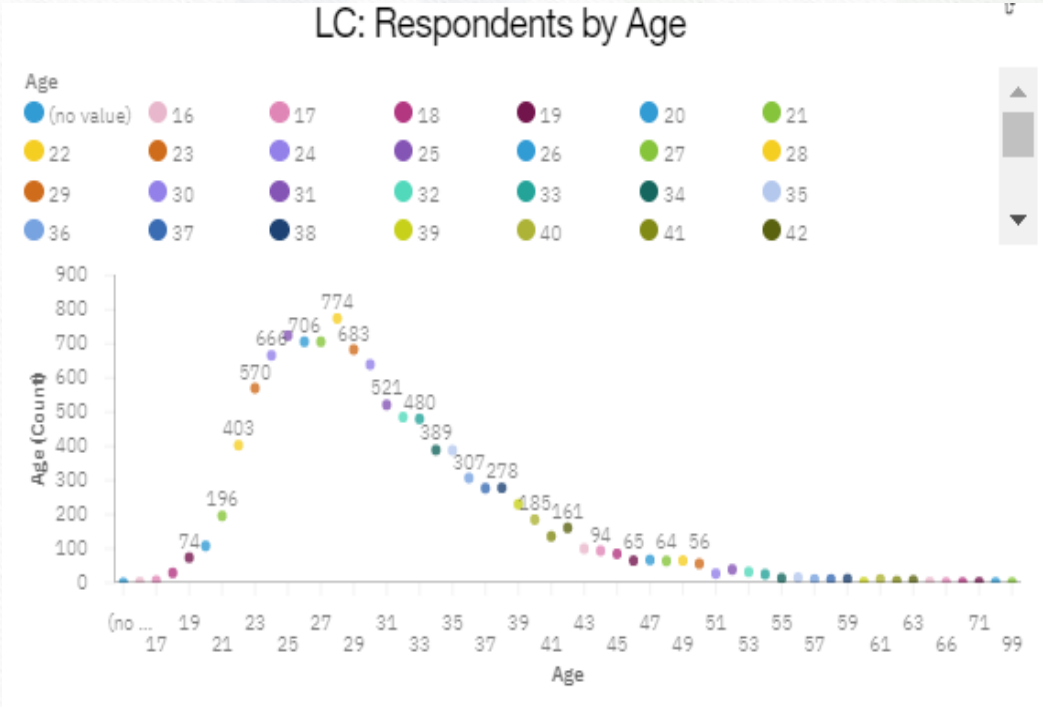

```
# your code goes here

lowSum = (dfNew['ConvertedComp'] <= (q1 - 1.5 * IQR)).sum()
print('Number of Outliers below Q1 is {}'.format(lowSum))
--------
uppSum = (dfNew['ConvertedComp'] >= (q3 + 1.5 * IQR)).sum()
print('Numberr of Outliers above Q3 is {}'.format(uppSum))
--------
numOutlier = lowSum + uppSum    #= sum(a, b)
print('Total Outliers are {}'.format(numOutlier))
```

```
Number of Outliers below Q1 is 0
Numberr of Outliers above Q3 is 0
Total Outliers are 0
```

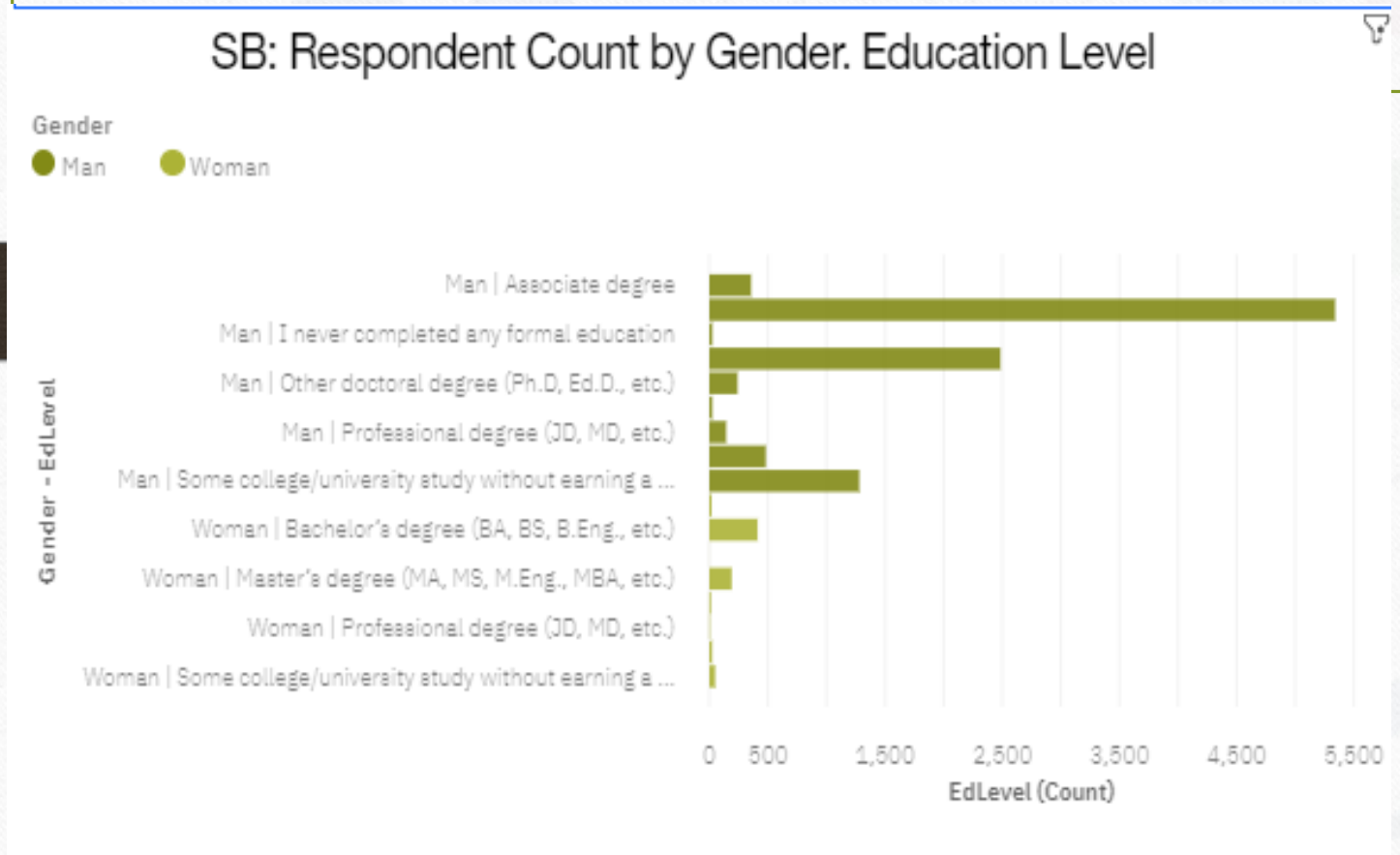
1. Inter-quartile range for salary is $65,984 – ~~single value around with 50% of data~~ points
2. Before removal of outliers, 933 were found beyond Q3 point of (193,976)
3. Median closer to Q3: distribution negatively skewed with mean < median

RESULTS: Visual Analysis



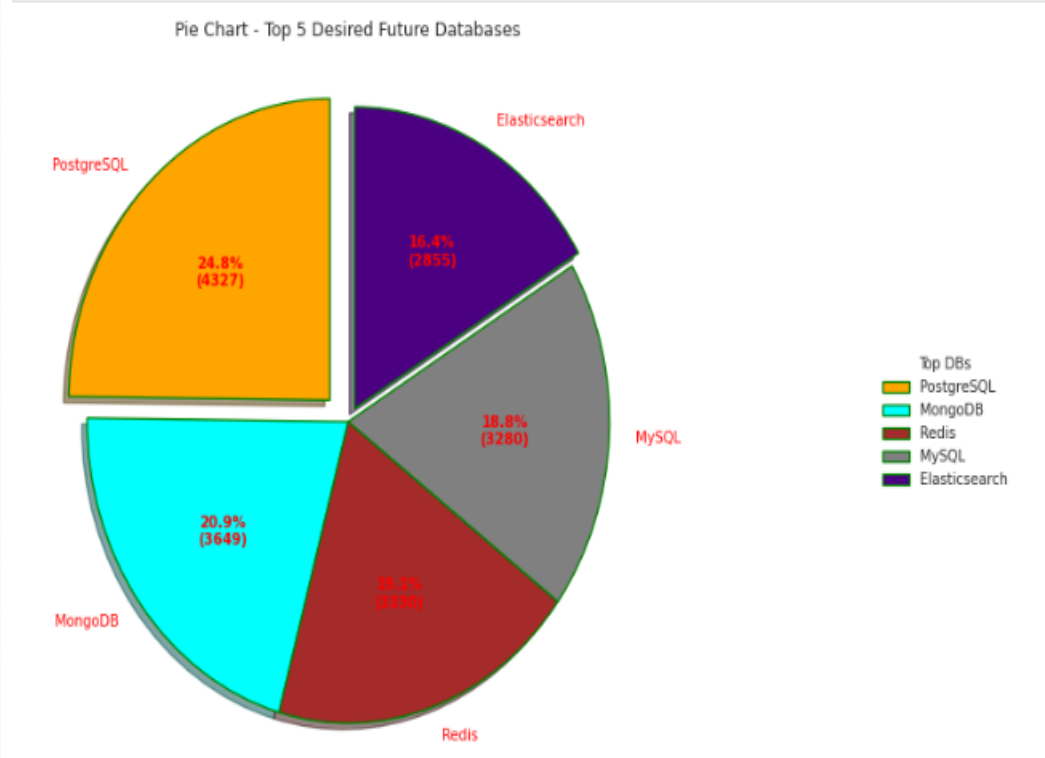
1. Respondent distribution by age follows normal bell shaped curve
2. Higher respondents in age range 19 - 40 years
3. Figure supports fact that median age of techs is 29 years

RESULTS: Visual Analysis



1. Men overall, shown as more qualified as more surveyed by far; difficult to draw sealed conclusion
2. Leading qualification for men with nearly 6,000 BA and nearly half as many Masters; while for women, BA of slightly < 1000 and Master around 500
3. More men with some college experience without a degree

RESULTS: Visual Analysis



1. PostgreSQL is future leader desired database (24.8%) followed closely behind with MongoDB at 20.9%
2. Represents an increase future desire for PostgreSQL from current 3rd spot use
3. Redis & Elasticsearch, the 2 lowest ranked here (as future desires) are not currently used;

RESULTS: Computational Analysis

Language	Avg. Annual Salary
Swift	$130,801
Python	$114,383
C++	$113,865
Javascript	$110,981
Java	$101,013
Go	$94,082
R	$92,037
C#	$88,726
SQL	$84,793
PHP	$84,727

1. Even with PHP recorded at lowest avg. annual salary, still a good salary showing in comparison to some fields

2. Average salary for Python, language of choice used in this analysis, is relatively competitive

3. Swift, like Python is a general purpose, open-source language tops chart with highest average annual salary $130,801

4. From scrapped data, avg median salary of $97,547; from survey, value of $57,745

RESULTS: Computation Analysis

2. Statistical/Numerical

   a. Median: AGE = 29.0 years: Outliers: *ConvertedComp* variable had 933 values above $193,976; after removing outliers, Mean salary value = $58,992.50

   b. Inter-quartile converted salary range:$65,984 – spread of 50% of data values

   c. *ConvertedComp*: Data Range: (-69,960, 193,976)

3. Relational Analysis

   a. Correlations: *CodeRevHrs* has –ve correlation of 0.01 with Age; As Age increases 1 unit, *CodeRevHrs* decreases by 0.01

   b. As Age increased a unit, *ConvertedComp* has slight +ve rise of 0.38 – not a particularly strong effect (below 0.50 mark)

RESULTS: Computation Analysis

1. Man & Woman with equal median salary of $57,745 (USD)
2. Gender seemingly not an issue: non-binary/genderqueer has highest median salary of $66,000

	Gender	ConvertedComp
0	Man	57745.0
1	Man;Non-binary, genderqueer, or gender non-con...	59520.0
2	Non-binary, genderqueer, or gender non-conforming	66000.0
3	Woman	57745.0
4	Woman;Man	25887.0
5	Woman;Man;Non-binary, genderqueer, or gender n...	30244.0
6	Woman;Non-binary, genderqueer, or gender non-c...	65535.5

Results: Pre-Imputation

	Respondent	CompTotal	ConvertedComp	WorkWeekHrs	CodeRevHrs	Age
count	11398.000000	1.058900e+04	1.139800e+04	11276.000000	8972.000000	11111.000000
mean	12490.392437	7.570477e+05	1.263096e+05	42.064606	4.781071	30.778895
std	7235.461999	9.705598e+06	2.846750e+05	24.672741	4.567060	7.393686
min	4.000000	0.000000e+00	0.000000e+00	3.000000	0.000000	16.000000
25%	6264.250000	2.500000e+04	2.901600e+04	40.000000	2.000000	25.000000
50%	12484.000000	6.500000e+04	5.774500e+04	40.000000	4.000000	29.000000
75%	18784.750000	1.200000e+05	9.500000e+04	43.000000	5.000000	35.000000
max	25142.000000	7.000000e+08	2.000000e+06	1012.000000	99.000000	99.000000

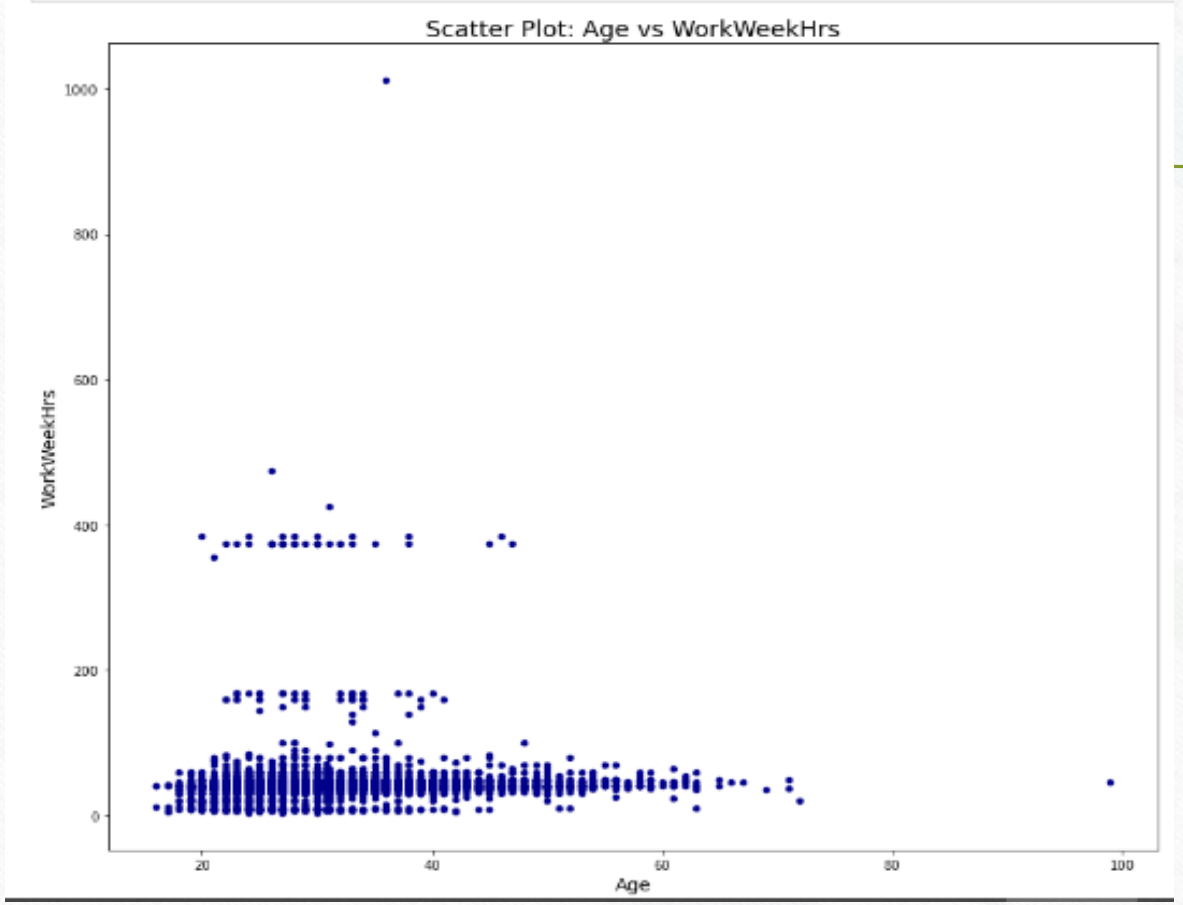
1. Count of non-missing values in all numerical fields are unequal
2. Mean(*ConvertedComp*)=$126,309
3. Mean(*Age*) is 30.78 years

Results: Post-Imputation

	Respondent	CompTotal	ConvertedComp	WorkWeekHrs	CodeRevHrs	Age
count	11398.000000	1.139800e+04	1.139800e+04	11398.000000	11398.000000	11398.000000
mean	12490.392437	7.570477e+05	1.263096e+05	42.064606	4.781071	30.778895
std	7235.461999	9.354789e+06	2.846750e+05	24.540330	4.051929	7.299999
min	4.000000	0.000000e+00	0.000000e+00	3.000000	0.000000	16.000000
25%	6264.250000	2.800000e+04	2.901600e+04	40.000000	2.000000	26.000000
50%	12484.000000	7.000000e+04	5.774500e+04	40.000000	4.781071	29.000000
75%	18784.750000	1.400000e+05	9.500000e+04	42.500000	5.000000	34.000000
max	25142.000000	7.000000e+08	2.000000e+06	1012.000000	99.000000	99.000000

1. Count of filled values in all numerical fields – 11,398 observations
2. Mean(*ConvertedComp*)= $126,309
3. Mean(Age) is 30.78 years - unchanged

RESULTS: Relational Analysis



1. Difficult to conclude exact correlation since eyeball fit of regression line would show a non-steep incline (flat)

2. Generally, correlation value of < 0.5 is weak and of no effect

3. All respondents favour longer work weekly hours (0.036) and less code reviews (-0.01)

4. Working hours of respondents [20 – 40] years range to 400 hours, max; while hours for techs 50+ ranges up to 100 hours.

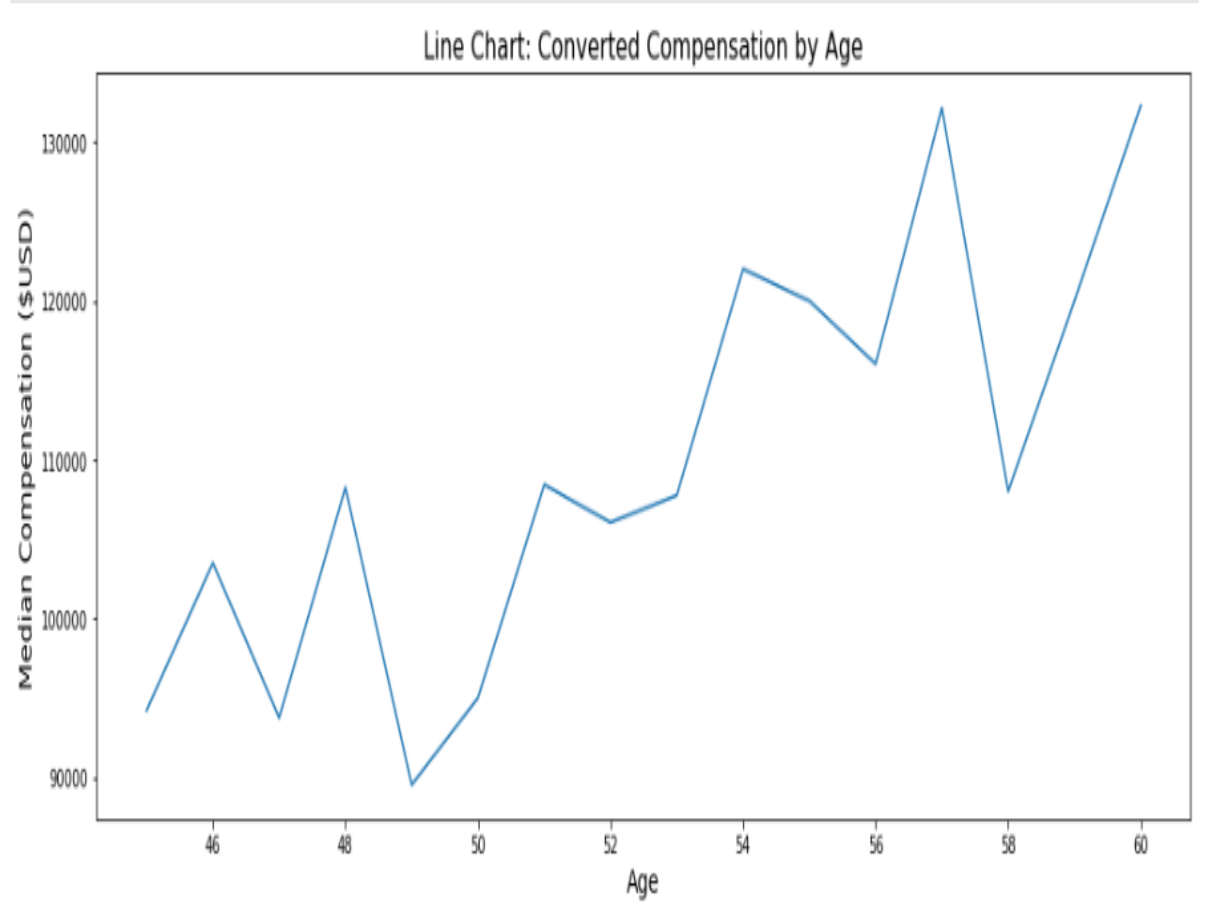
RESULTS: Relational Analysis – Correlation

```
df.corr()['Age']
```

```
Respondent        0.003991
CompTotal         0.006732
ConvertedComp     0.101822
WorkWeekHrs       0.036311
CodeRevHrs       -0.017042
Age               1.000000
Name: Age, dtype: float64
```

1. No single variable has strong correlation (+ve/-ve) with Age
2. ConvertedComp has highest, value of 0.10

RESULTS: Trend Analysis – Line Chart



1. Overall Converted salary has increasing trend with age overtime.

2. Lowest median salary recorded for 45 year old is $90,000; highest median salary of over $130,00 for techs 60 years .

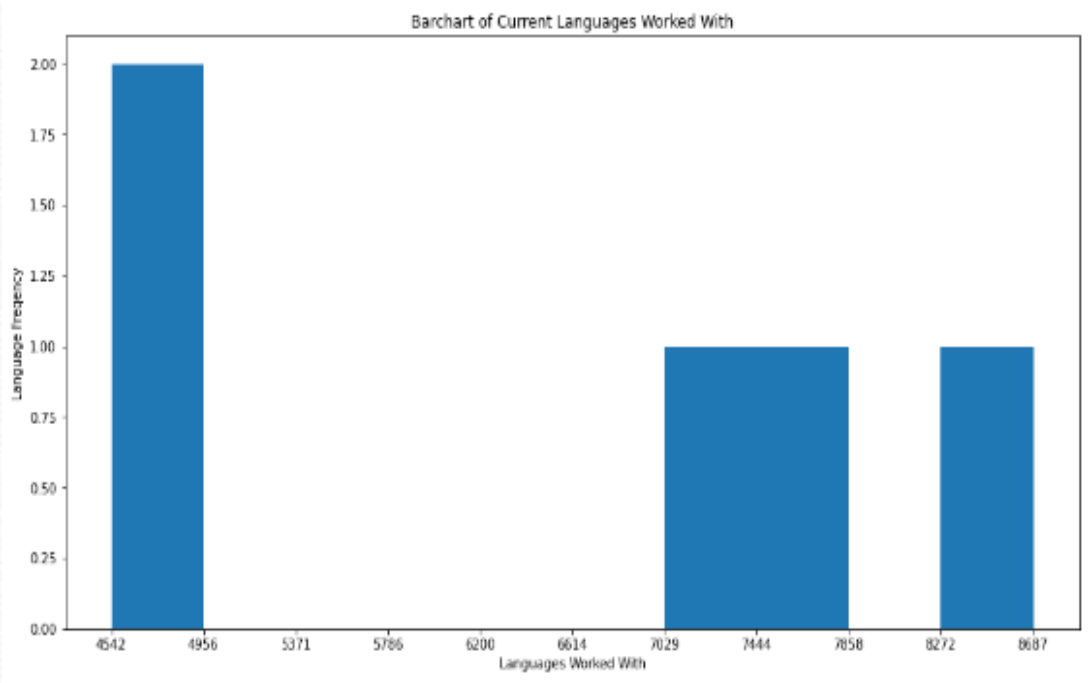
3. Similar salary earnings between $(105,000, 110,000) for persons between 52 and 53 years

4. Highest salary of approx. $(110,000, 132,000) for persons between 58 and 60 years

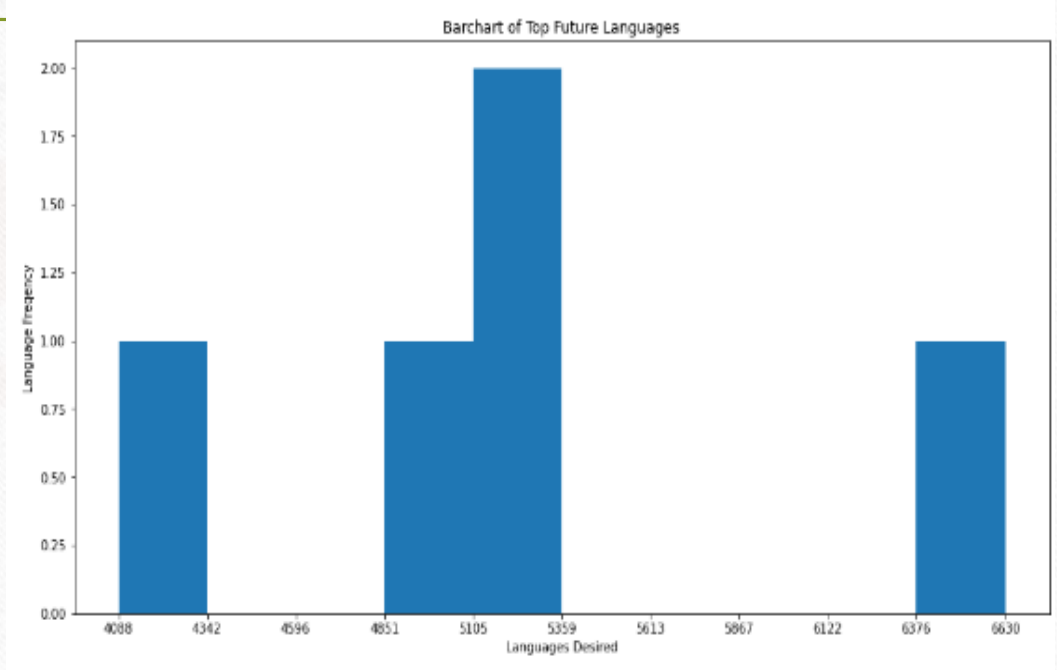
PROGRAMMING LANGUAGE TRENDS: top 5

Language

Current Year



Next Year



PROGRAMMING LANGUAGE TRENDS -

FINDINGS & IMPLICATIONS

Findings

1. Twice as many respondent who worked with given language will favour working with languages: range to increase from [4524, 4956] to [4551, 5359] in future
2. Alternatively, the frequency of language count at level 1 will decrease from 8757 to 6639; Decrease in demand of current top 3: JavaScript, Html & SQL
3. Current 3 top ranking languages (JS, HTML, Sql) will all see decrease future usage

Implications

1. Generally, techs are in favour of acquiring more language skillsets; geared more towards object-oriented, open source, web-based and easier readability (English-like)
2. Number of techs favouring a single language skillset will decrease
3. Wide spread -these represent different concepts – frontend, web design and query language resp . Replacements for these will favour more object-oriented coding, easier connectivity and compatability, simplier syntax.

PROGRAMMING LANGUAGE TRENDS -

FINDINGS & IMPLICATIONS

Findings

4. Expressed increase use for Python from 4542 to 5239 by almost 800 subscribers

5. Scripting language replacement of Bash/Shell/Powershell with Typescript

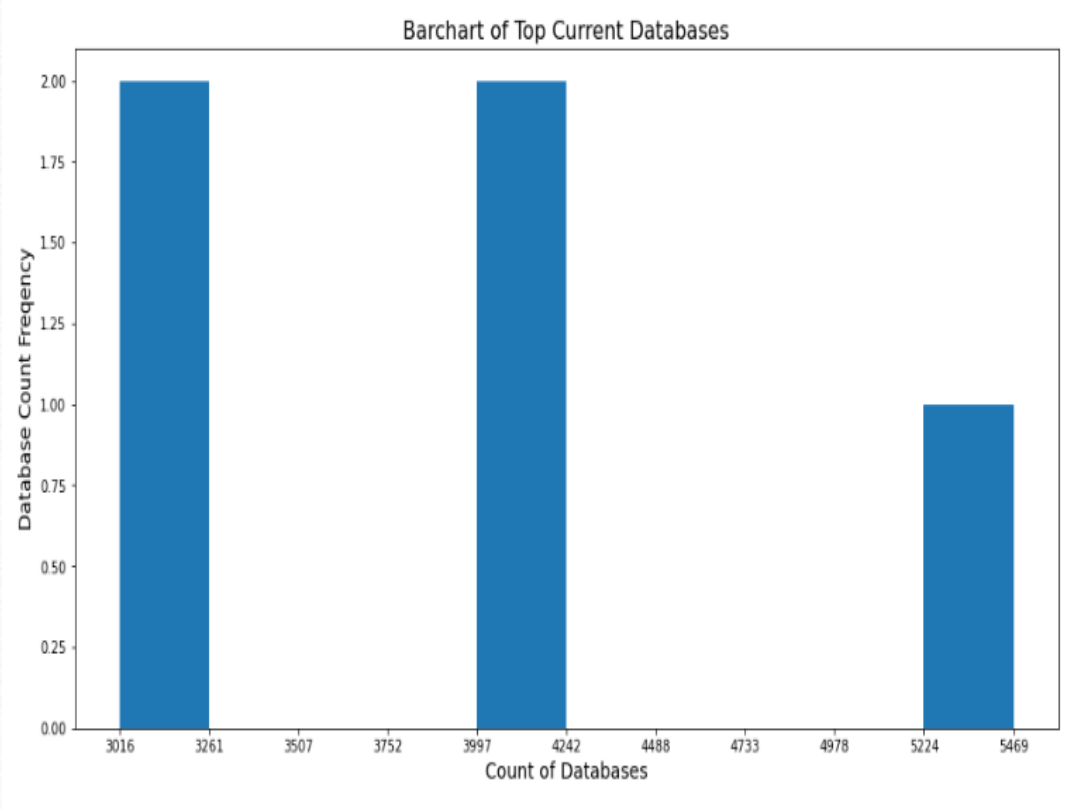
Implications

4. Python – free, open-source, portable, dynamically typed, English-like posed to become among the leading languages; continued strengthening of its community; wider use even by younger grouping worldwide eg primary schools

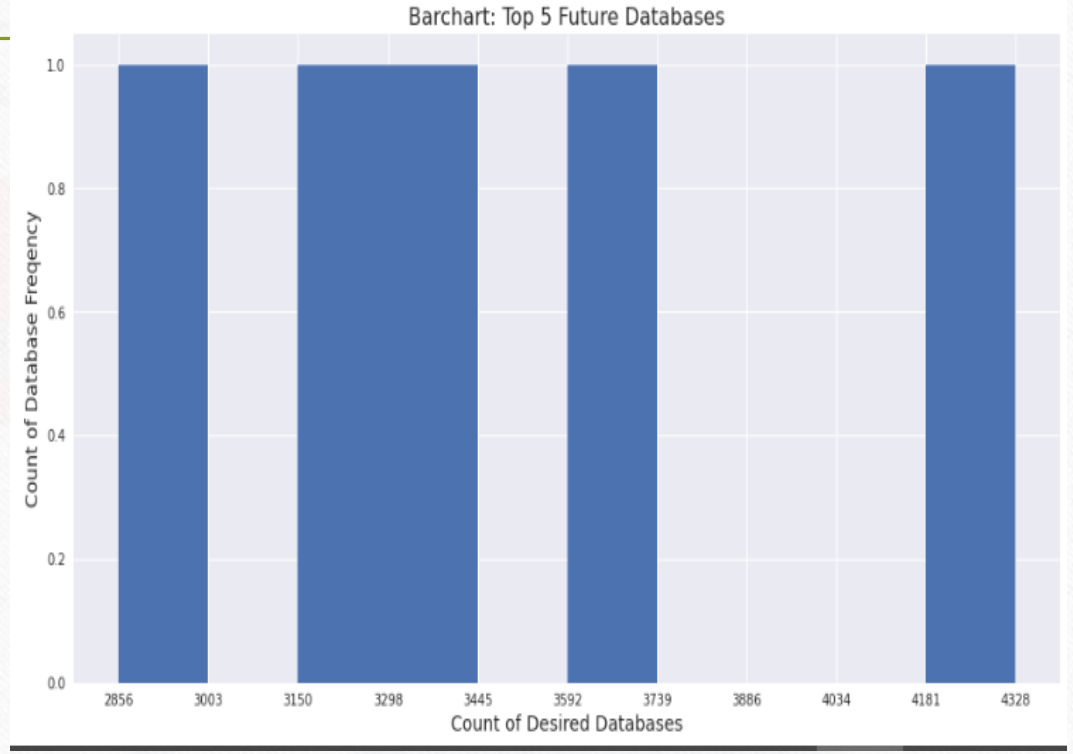
5. Desire to use open-source object-oriented interface to replace server-side scripting; embrace of new technology and possible willingness to maintain old ones

DATABASE TRENDS: top 5 Databases

Current Year



Next Year



DATABASE TRENDS - FINDINGS & IMPLICATIONs

Findings

1. Maximum number of database of frequency count of 2 will increase from a max value of 4242 to max value 5359
2. Overall distribution of current count has tapering right tail, while shape of future distribution resembles a uniform distribution
3. Use of Postgresql, SQLite & MongoDB will remain constant

Implications

1. Higher number of respondents favor greater number of database use; increase database usage will broaden field; companies will have options; Challenge could be hardware compatability , data transfer and data communications which may cause these areas to see growth
2. Currently at level 1, max databases used is 5469 but in future projection is 4328; Changing field for database technology on the rise favouring higher standard
3. Database developers should want to tap into these by increasing knowledgebase and skillsets; systems not employing use may want to revert; If constant trend is continued as more developers may want to access skills in these 3 dbases

DATABASE TRENDS - FINDINGS & IMPLICATIONs

Findings

4. Future preference for new databases: Redis and Elasticsearch

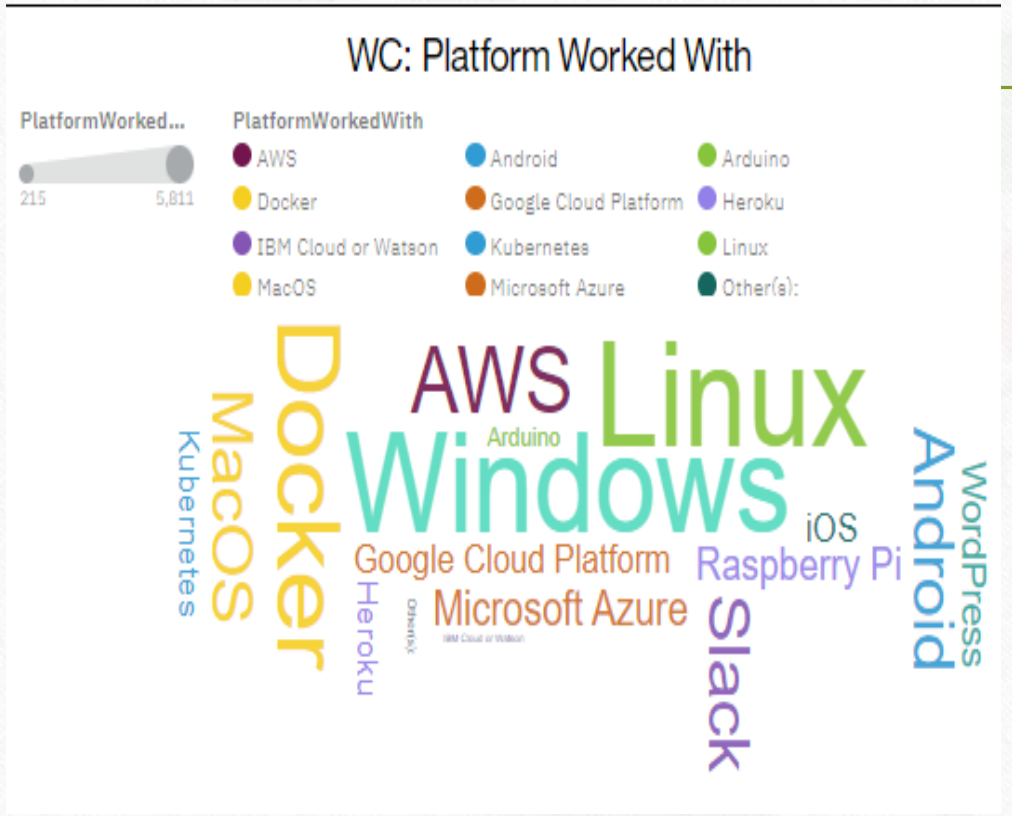
5. Current strong desire that MS Sqlserver will become obsolete

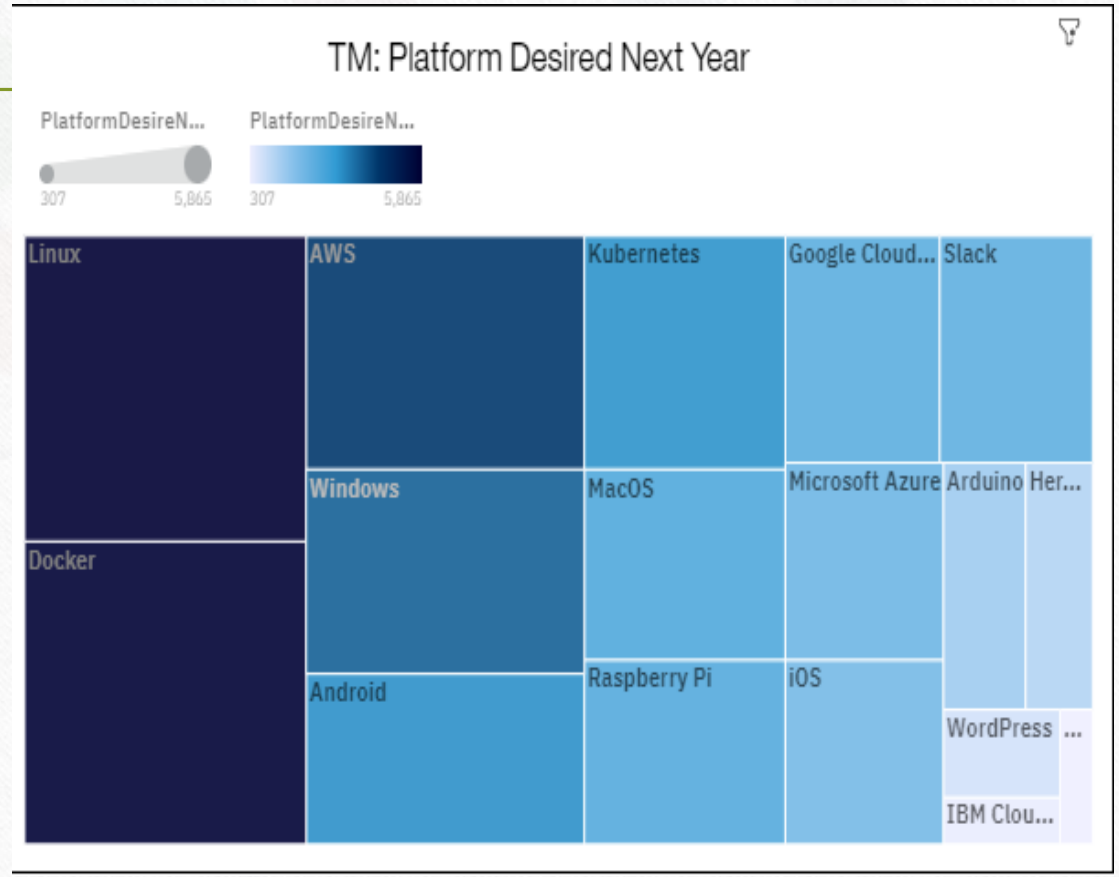
Implications

4. Organizations may faced dilemma of changes to comply with newer preferences – adjusting data landscape to keep up with newer demand or find enough techs to maintain current ones;

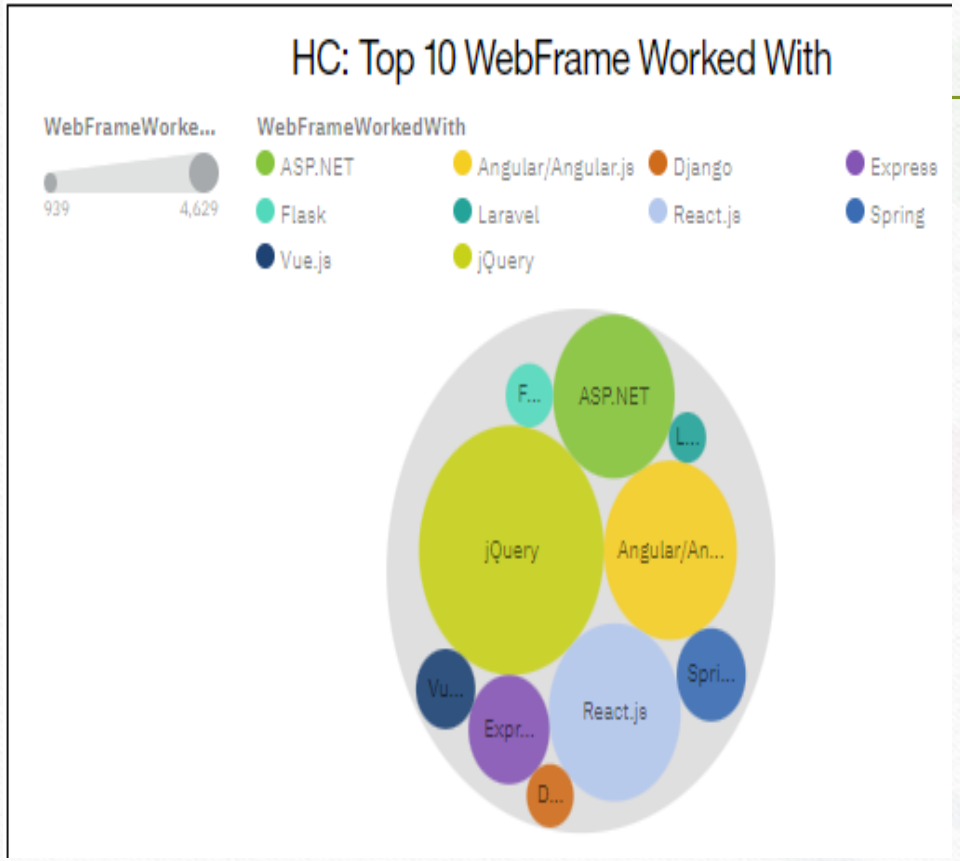
5. Systems built on these platforms will be force to upgrade, or struggle to maintain skillset for support; already acquired skills will need to be upgraded

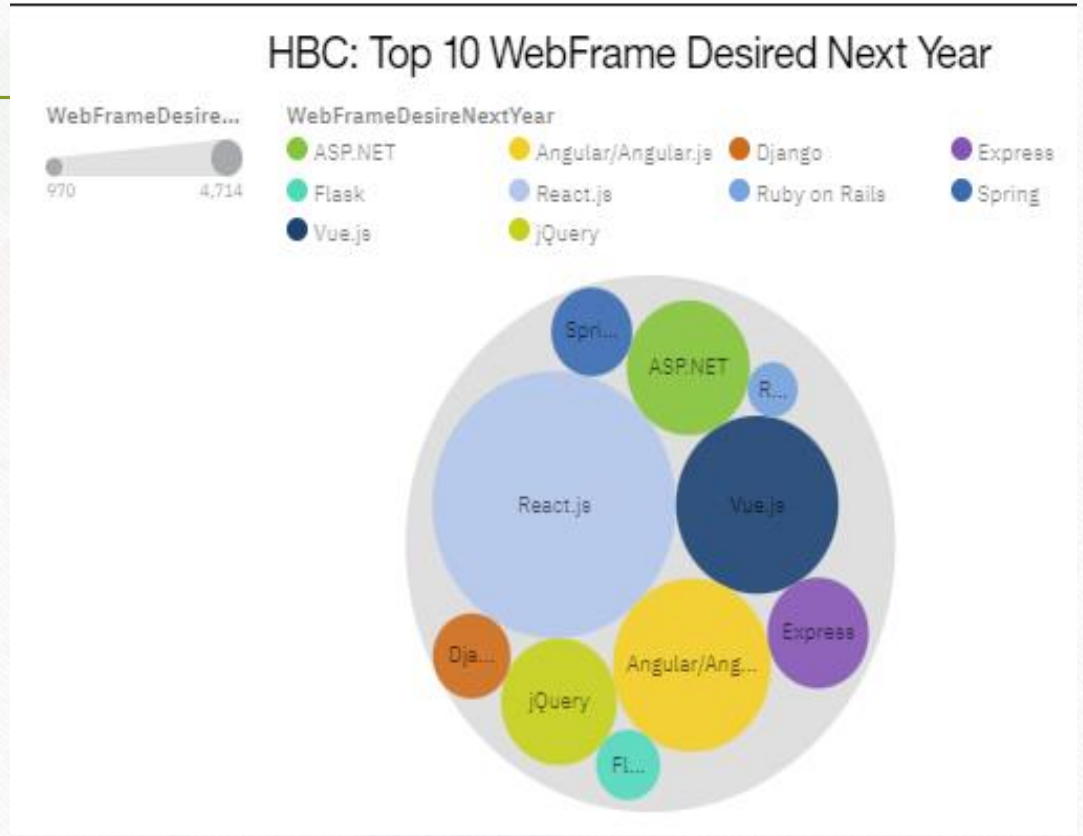
TRENDS IN PLATFORMS – visual comparison





TRENDS IN WEBFRAMES - visual comparison



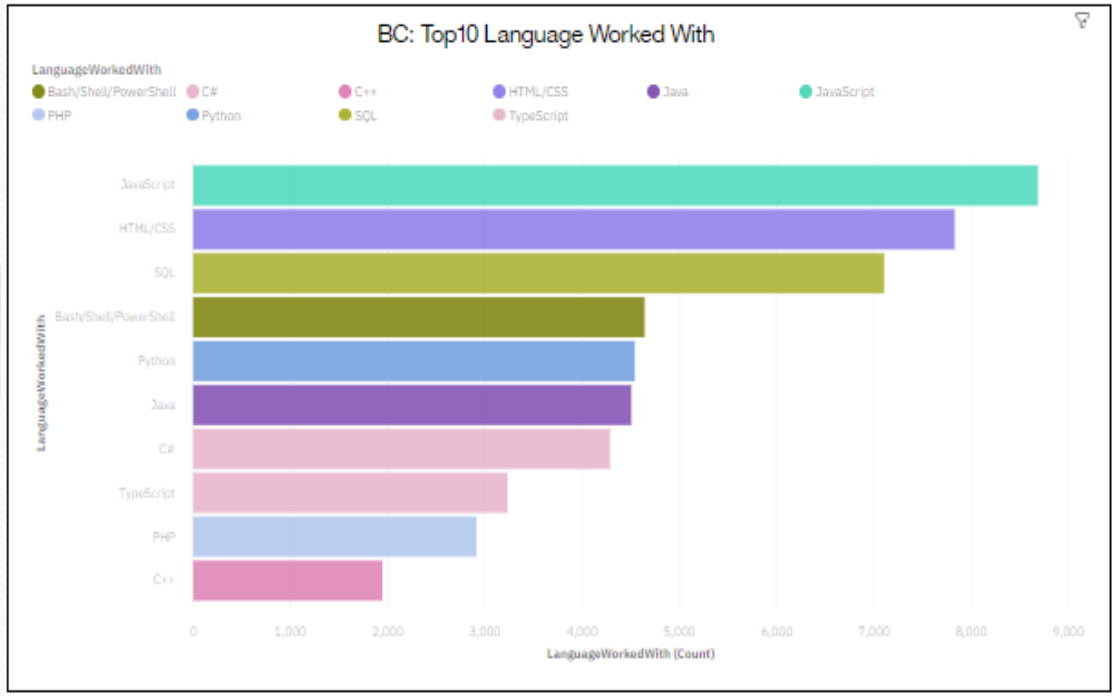


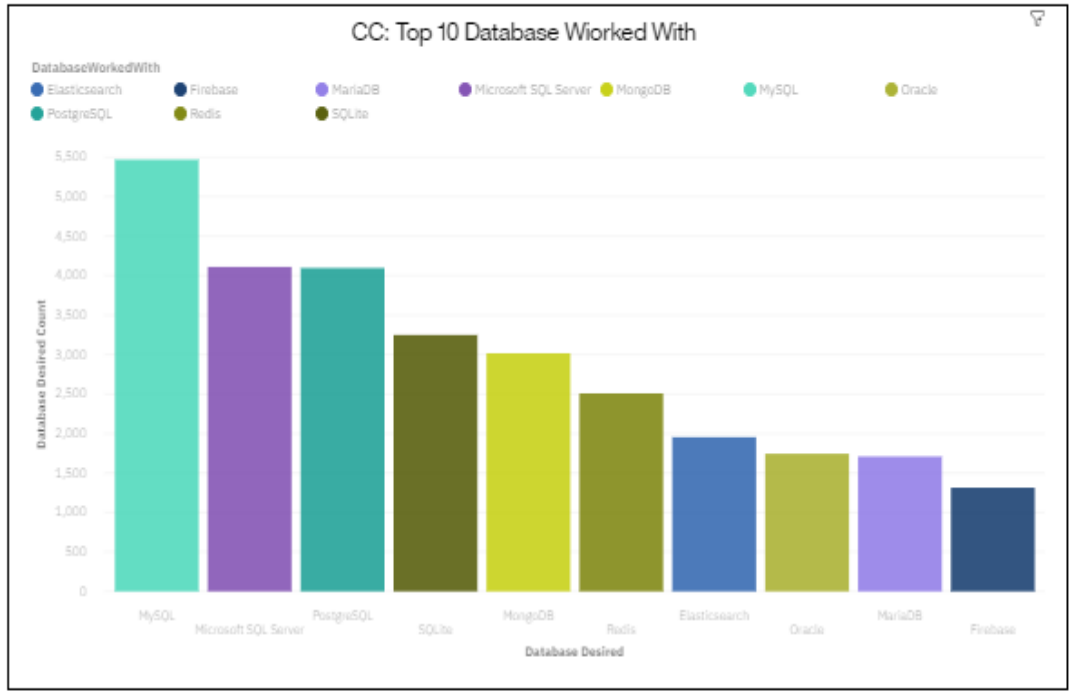
DASHBOARD Link

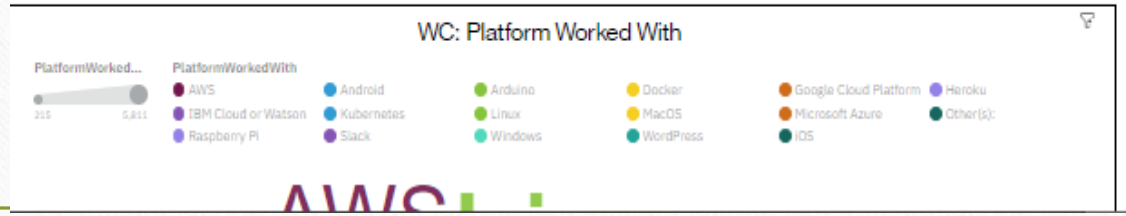
https://dataplatform.cloud.ibm.com/dashboards/6ee20e50-3d03-4e40-b374-c6c2d7d6159d/view/4268c418619369d043afb1e4079b7d0f7b3f2059e0bb830b84817b495e667697f0681394c8274a09de440c35f0ef410dcc

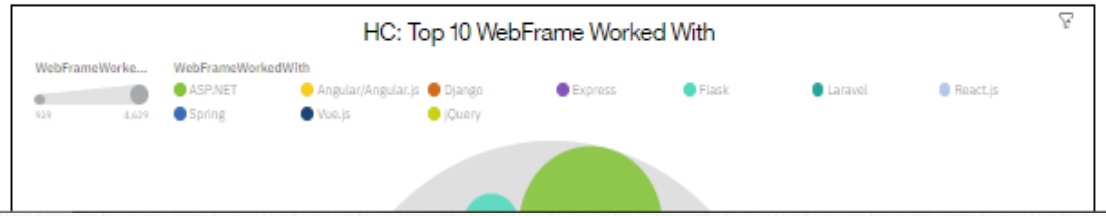
DASHBOARD 1: Current Technology Usage

Current Technology Usage | Future Technology | Demographics



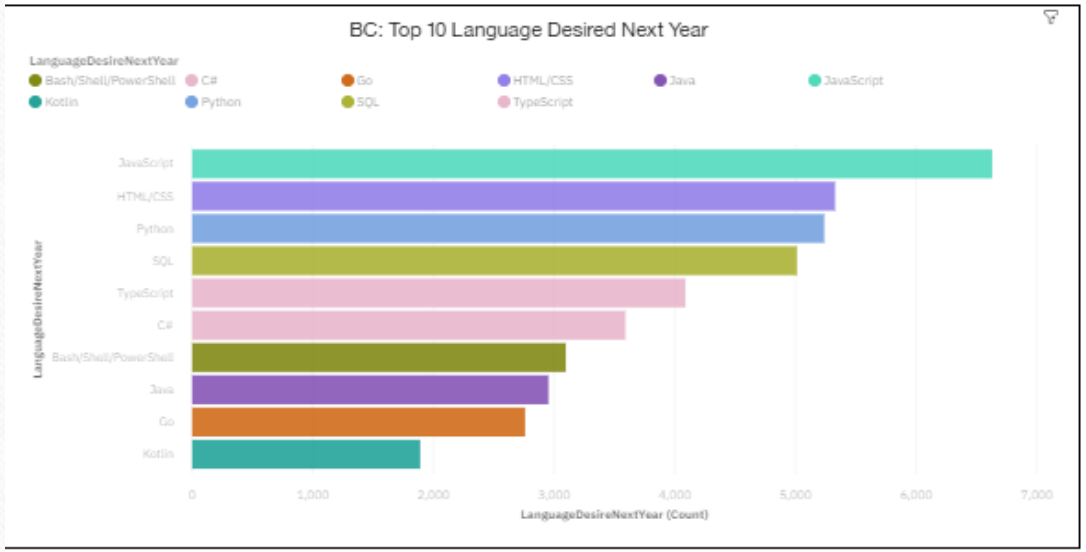


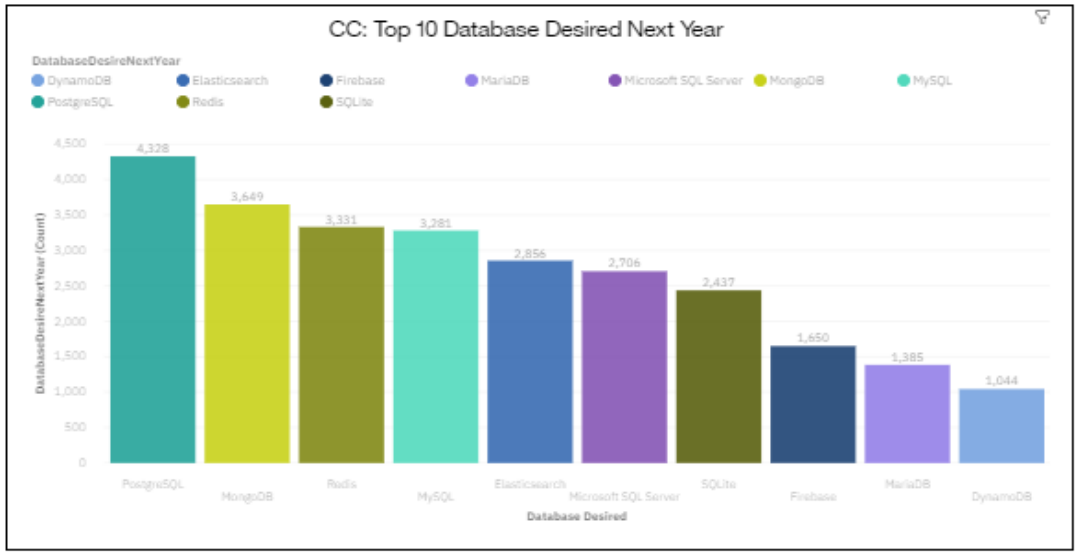


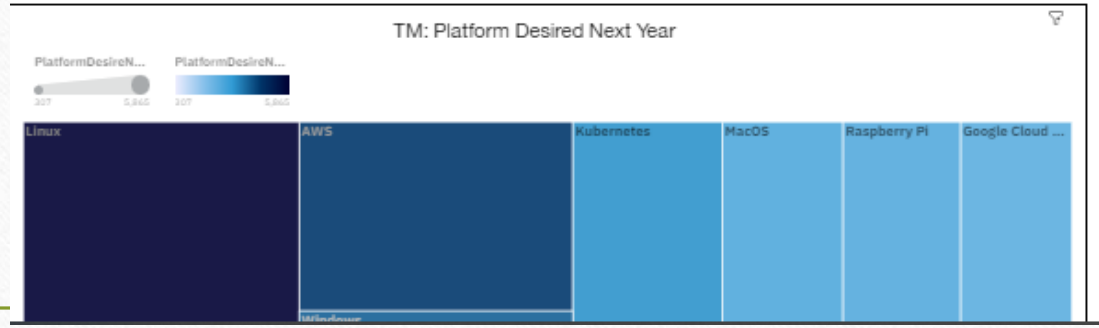


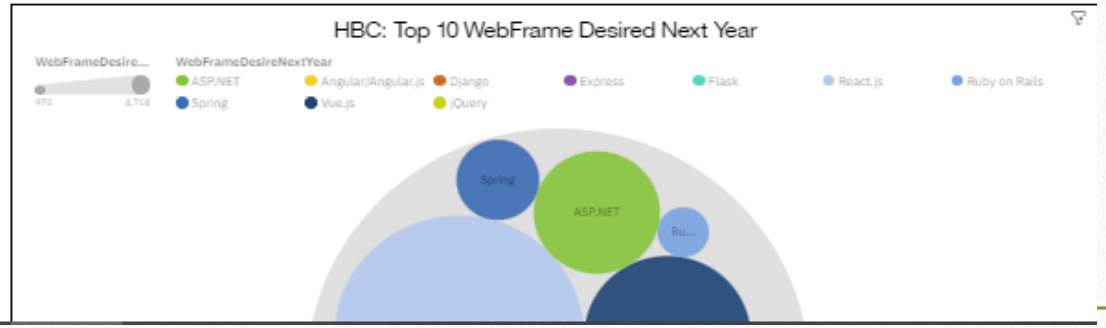
DASHBOARD 2: Future Technology Usage

Current Technology Usage | Future Technology | Demographics



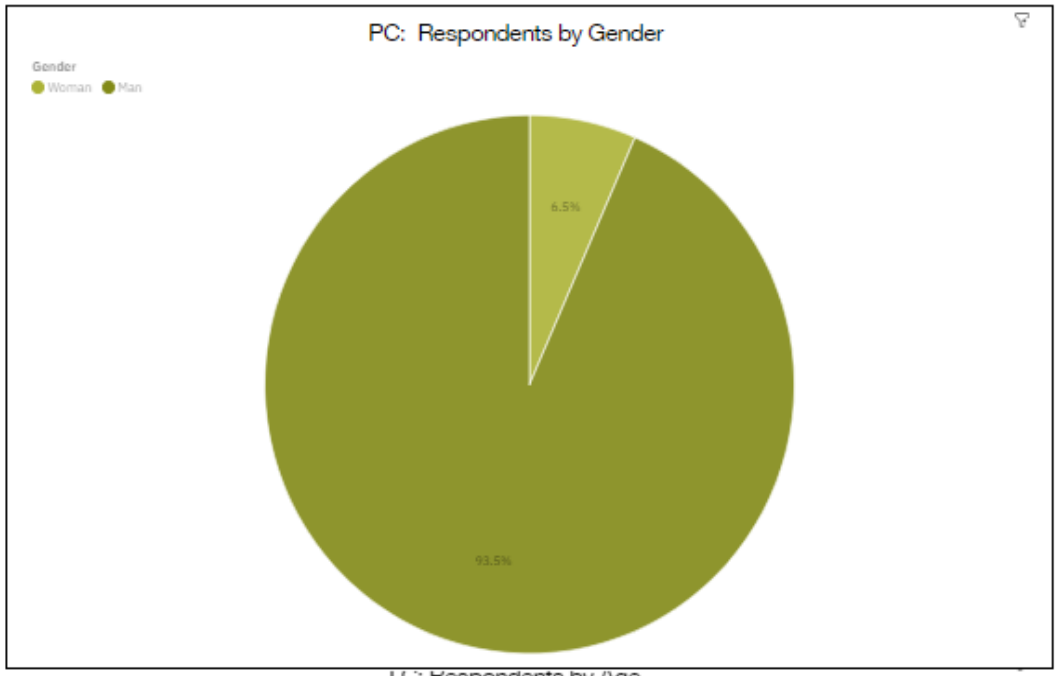


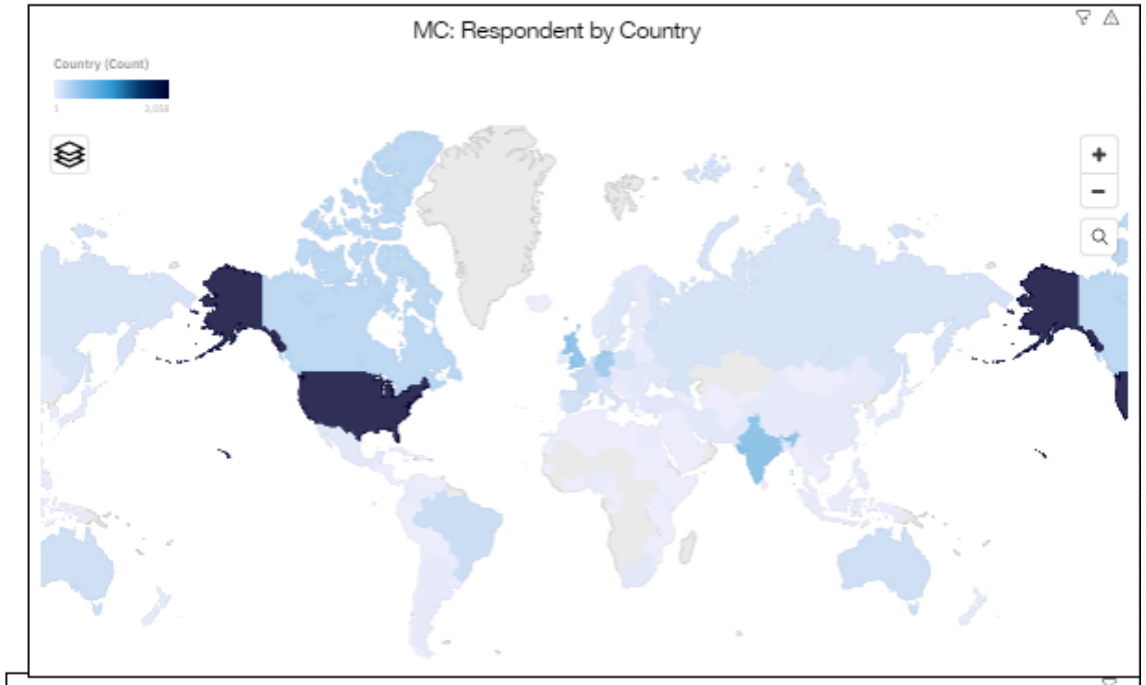


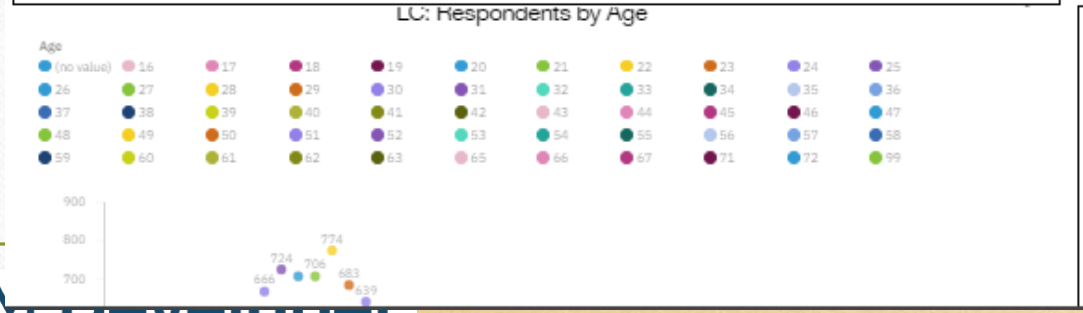


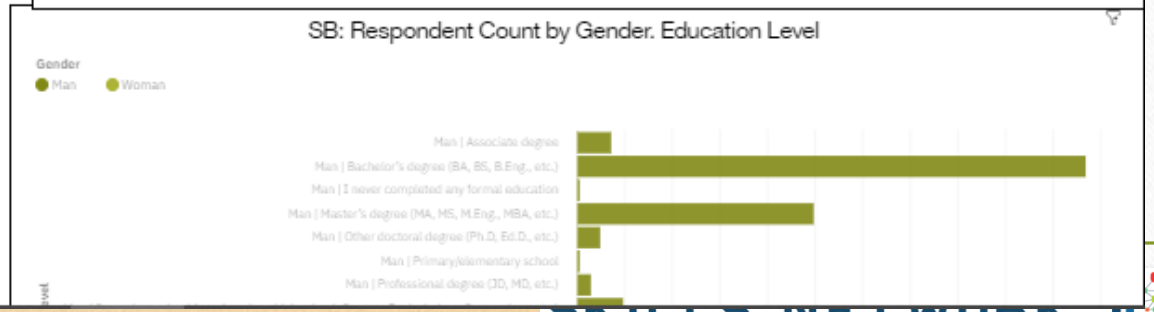
DASHBOARD 3: Demographics

Current Technology Usage | Future Technology | Demographics









IBM Developer
SKILLS NETWORK

COURSE LINKS - GitHub

1. Collecting Data using WebScraping - https://github.com/bahamalytixs/Capstone-Module/blob/main/HR-Web-Scraping-Lab-Final.ipynb
2. Collecting Jobs Data: https://github.com/bahamalytixs/Capstone-Module/blob/main/HR_Collecting_Jobs_data_Using_API-Questions-Final%20(2).ipynb
3. Exploratory Data - https://github.com/bahamalytixs/Capstone-Module/blob/main/HR-M3ExploratoryDataAnalysis-lab-Final.ipynb
4. Data Visualization - https://github.com/bahamalytixs/Capstone-Module/blob/main/HR-M4DataVisualization-lab-Final.ipynb

DISCUSSION: Analytical Findings

1. Duplicate records: Removal of these provide more data integrity since survey would want single response each participant
2. Imputed values: computational analysis changed when all data values filled with representative values
3. Normalization: Data results made more sense of so that no single value would have higher influence on output eg weighing of *compfreq vs totalcomp*
4. Salary distribution: Observed long tail, signaling many outlying data points. Can conclude that a number of IT personnel possibly engaged in contractual/self-employed work
5. Relational Analysis: In particular no numerical value has significant correlation with age; converted salary variable has highest impact of 0.38. While a low *p-value* was noted, can only conclude that the slight r value of 0.1057 could be attributed to experience

DISCUSSION: Computation & Trend Analysis

1. Computation Analysis
   a. Median Age is 29 years of responders
   b. ~~No gender discrepancy with remuneration; both genders get equal pay of $57,745(US~~ dollars) even though males more qualified
   c. Number of male respondent outnumbered female – 10,480 to 731. Implication that more males in technology field. Future exploration which fields specifically?
   d. Non-binary, genderqueer had highest salary of $66,000

2. Trend Analysis

   a. Overall Converted salary has increasing trend with age, overtime

   b. Lowest median salary recorded for 45 year old is $90,000; highest salary of over $130,00 for techs 60 years.

   c. Similar salary earnings between $(105,000, 110,000) for persons between 52 and 53 years

   d. Highest salary of approx. $(110,000, 132,000) for persons between 58 and 60 years

DISCUSSION: Prog. Language Analysis

3. Language Analysis

   a. Generally, techs are in favour of increasing access to different languages

   b. Decrease in current top languages - Wide spread effect representing different concepts – frontend, web development and querying languages; Respective communities should come together and address the decline of these leakages;

   c. Python expressed as increase use. May trigger coding world in this direction

   d. Typescripting – Coders should embrace this so that systems using other scripting languages could be up-graded or remain current to maintain current systems using Bash/Shellscript;

DISCUSSION: Database Analysis

4. Database Analysis

   a. Maximum number of database of frequency count of 2 will increase from the max value of 4242 to 5359 indicating that a higher number of respondents favor access to greater number of databases

   b. Overtime, more developed technologies will have count number falling in range [4596, 5867] to shape distribution as normal curve

   c. Expressed consistent use of Postgresql, SqLite & MongoDB: since these cited for future use, developers would want to continue and remain relevant with these; further, the community support should continue;

   d. Future preference for Redis & Elasticsearch given that these have 0 current usage. New area for developers to launch into.

   e. No cited future preference for MS Sqlserver; systems built on these platforms will have to pivot to remain relevant or pay price for future support

DISCUSSION: Platform Trends

5. Current Platforms

   a. Higher platform preferences noted by larger word-size: Windows, Linux, Docker

6. Future Platforms

   a. Seeming declining preference for Windows as Linux and Docker has higher ratings

   b. Introduction of Slack, a messaging app for businesses that connect people to needed information, as future technology

   c. Arduino and Heroku will be in competition

DISCUSSION: WebFrame Trends

7. Current WebFrames
    a. Current leaders: jQuery, React.js, ASP.net & Angular as noted by bubble size
    b. From eyeball view, React.js, ASP & Angular are competitors

8. Future WebFrames
    a. Cited as future constants: ASP and Angular
    b. React.js now poised to become leader with decline of jQuery
    c. Vue.js webframe to show stronger presence

OVERALL FINDINGS & IMPLICATIONS

Findings

1. IT Salary Distribution:

2. US IT based opportunities by openings & location

Implications

1. Observed value from survey, converted salary has long right tail signifying potential outliers; Q3 value is $ 193,976 ; Great number of self-employed setting their costs thru perhaps consultancy;

1. IT job posting opportunities implies that there is greater potential for growth; To compare study needs to consider other industralized countries. Generally, not a good showing in the IT field; Is the field saturated? No more IT jobs? Are all businesses using technology to the max? All questions based on recorded figures. For given openings, relocation is needed to colder states; Opportunity for southern cities to revisit IT landscape and create more opportunities - encourage more IT based projects;

2. Great demand for C language skills by far but no comparative salary noted for the same

OVERALL FINDINGS & IMPLICATIONS

Findings

3. How IT personnel remunerated

Implications

1. 1. Scrapped salaries range from $84,727 to $130,801; with median salary of $97,547. From survey, median salary is $57,745 after data cleansing; Is it that persons are more reluctant to discuss salaries when questioned upfront? Does the true value lie some where between?

2. 2. Both male and females rewarded the same - median salary = $57,745, despite males having higher qualifications; Based on global trends, remuneration is adequate in comparison; Salary not gender-based from web dataset;

OVERALL FINDINGS & IMPLICATIONS

Findings

4. Specific characteristics for IT specialists eg. age, work-hrs, etc?

Implications

1. Age range from 16 – 99; with 29 being average age; Shape of age distribution favours younger persons with age range 16 - 45
2. Generally, poor correlation between age and other numerical variables (*WorkWeekHrs, CodeRevHrs*); Other variables may prove more significant
3. WorkWeekHrs showed weak correlation of 0.036 summarizing that as age increases, total week work hours has little effect. Either way, all techs work longer weekly hours; both show commitment to working longer; younger ones (<50 yrs) spend longer hours reviewing codes; Older persons spend both less time working longer hours and reviewing codes perhaps by nature of their age and experience to do other tasks (system design, dbase management)
4. CodeRevHrs has –ve correlation of -0.01: at higher age, less time spent reviewing codes; Implication could be that older persons prefer other IT tasks like system development or consulting
5. From another view, all age groups prefer longer week hours and less code review hours; implication could be that techs occupy their time with other possible IT tasks – system design, coding or data modeling

OVERALL FINDINGS & IMPLICATIONS

Findings

Implications

1. Over 95% of surveyed are males signifying either male dominated field or survey bias selection
2. Educationally, men appears more qualified; higher number of advanced degrees compared to women: results may change if more females surveyed; consideration needed of location and timing of survey
3. Surprisingly, median average salary for both male and female are the same; additional qualification do not make difference
4. Younger persons in IT with median age of 29 years
5. Younger techs more dedicated to longer work hours: Is this greater commitment or passion to succeed or are they more driven?
6. Younger techs favour more code review hours: Do they need longer time because of learning curve?

OVERALL FINDINGS & IMPLICATIONS

Findings

5. Reactions of IT personnel to future technology?

Implications

1. Futuristically, preference to change current IT landmass: dbases, program languages,, webferames & platforms; speaks to changing mindset of younger generation to interact with easier to manipulate tools, across broader spectrum
2. Python is only currently used language with noted increase future use; Techs would want to remain current in this language; Community should want to continue to support knowledge base sharing; Encouragement for more programmers in this field
3. Top 3 language leaders (JS, HTML, SQL) will all have decline use; Opportunity for IT techs to remain current to service systems that will not welcome change and embrace newer skills.
4. Survey showed that database preference would remain stable moving into future; Currently used databases (Postgresql, SQLite, MongoDB) would be shuffled into top spots;Techs need to sharpen skills to remain relevant; Current users have decision to make in terms of database transition
5. Preference for new databases (Redis, Elasticsearch) also shown; Opportunity for techs to also embrace systems that house current databases noted to not advance to future

OVERALL FINDINGS & IMPLICATIONS

- *Reactions of IT personnel to future technology… cont.*

6. Slight declining preference for Windows as Linux and Docker has higher ratings: most unwelcomed effect since Windows is commonplace household favorite for years
7. Introduction of Slack, a messaging app for businesses that connect people to needed information would be welcome in IT project management since IT projects' successes are minimal; Further with sales being lowest IT function, Slack technology can play vital role as communicating tool
8. Uses of ASP.net & Angular cited to maintain constant presence in future: Speaks to phased approach to replacement of current websites and their ancillaries
9. React.js poised to become leader with decline of jQuery: Realized need for easier coding, dependencies management, template design, cross platform support

CONCLUSION: Findings

1. How are variables of the data typed & distributed?

2. What are some US based IT opportunities?

3. How are IT personnel remunerated?

1. Variables mainly typed as objects (string); Distribution shape captured by bar chart; Age, positively skewed; converted salary, negatively skewed (larger right tail)
2. Job openings in scattered cities; low counts in southern states; Given US as first world country, counts not particularly representative of leading economy; No citing for central cities; Technically, only C programmers in relatively high demand; Overall, avg annual salary relatively acceptable [84727, 1340,801]
3. Remuneration listed for a number of programming languages from $130,801 to $84,727 with no apparent gender bias; largest median salary quotes for genderqueer; differences in median salary shown in salaries scrapped from web and ones reported by survey

CONCLUSION: Findings

4. Are there specific characteristics for IT specialists eg. age, jobFactor, WorkLoc, etc?

4. At least 40 hours of work per week; lesser hours for code reviews for both young and older persons

5. What are sentiments of IT personnel on future technology - Languages, Database, Platform, WebFrame?

5. General preference for change in technology:

a. Language: Increase use of Python and introduction of TypeScript

b. Database: Among top 5 citings, continued use of 3 current dbases and intro of Redis & Elasticsearch

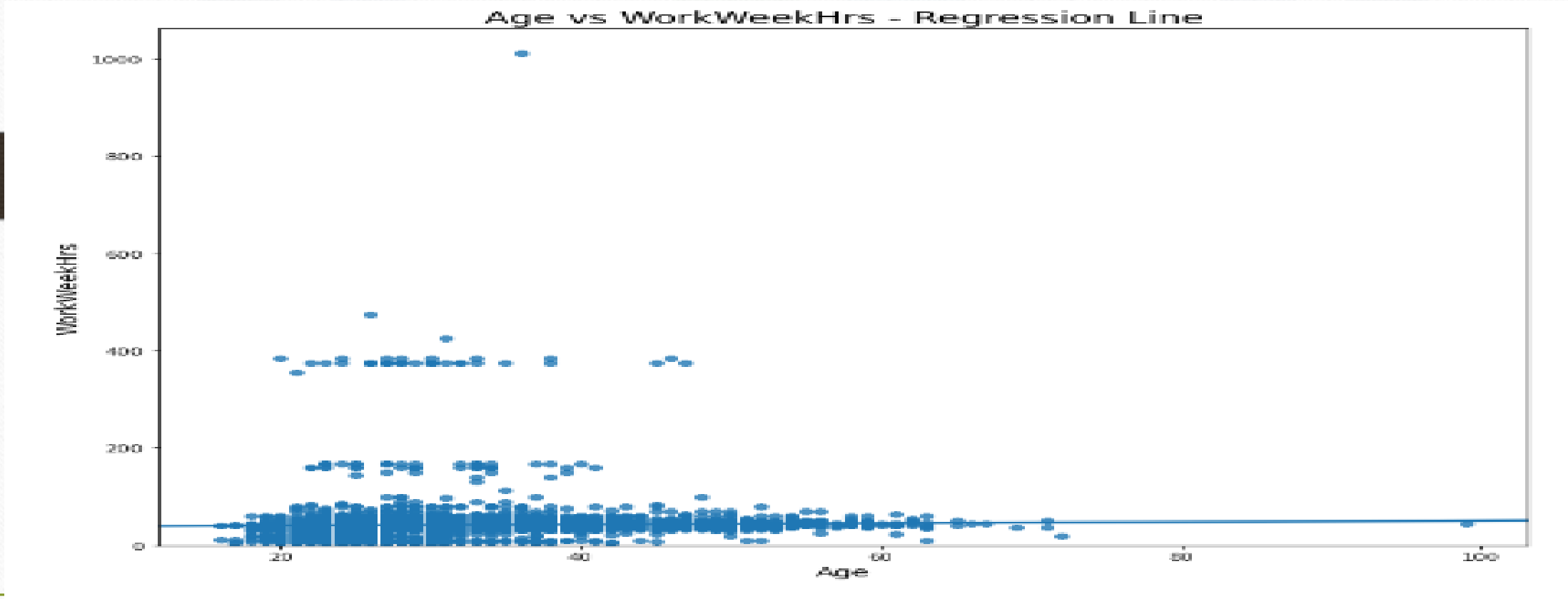
c. Platform: Declining preference for Windows, a household fixture with introduction of communicating app, Slack

d. Webframe; React.js, now ranked in top 3, voted to become future leading webframe; ASP.net & Angular are competitors and are to retain constant use in future

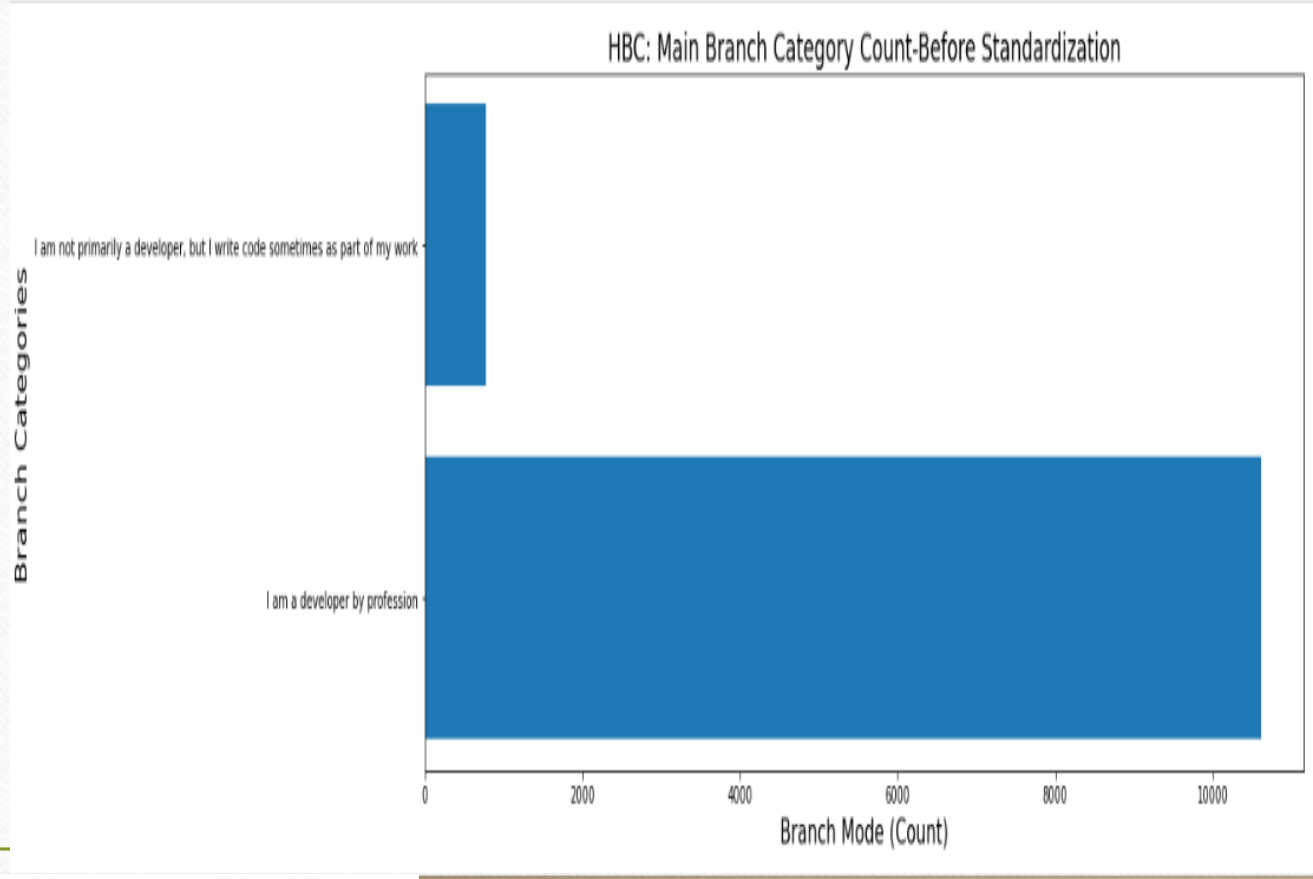
CONCLUSION: Future Analyses

JobFactor	What job factors are important to IT enthusiasts?
WorkChallenge	What are the greatest challenges in IT productivity
WorkLoc	Do IT techs have preference for work locations post pandemic?
SocialMedia	Do techs have preferred social media?
PurhaseWhat	What are techs level of influence regarding company purchase?
Impsum	How do they rate their competency based on work & experience?

APPENDIX: Correlation – Regression Line



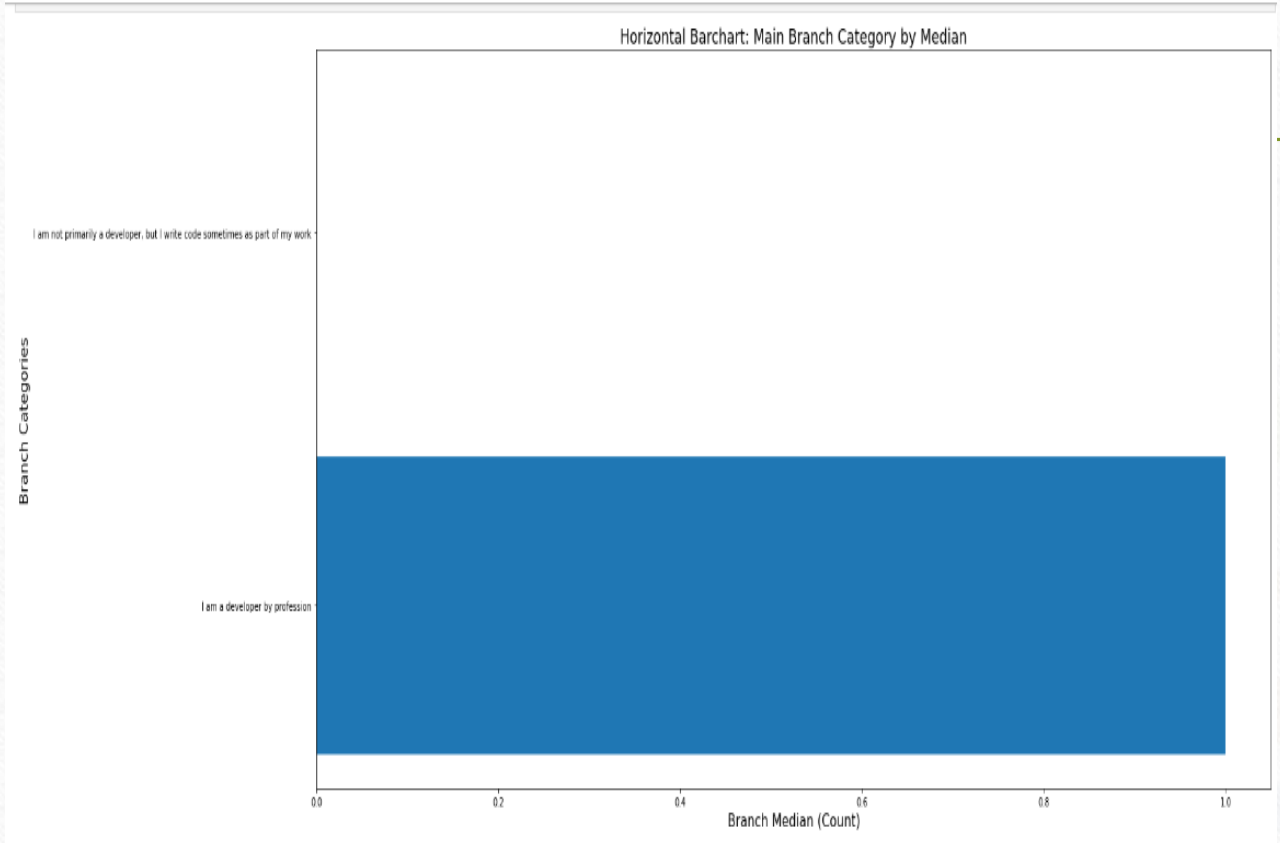
APPENDIX – Developer Analysis



Pre-normalization

1. Showing number of respondent who are developers vs part-time developers
2. True developers outnumber part-time ones by about 10:1
3. To make sense of analysis, need to recode and normalize data to balance values

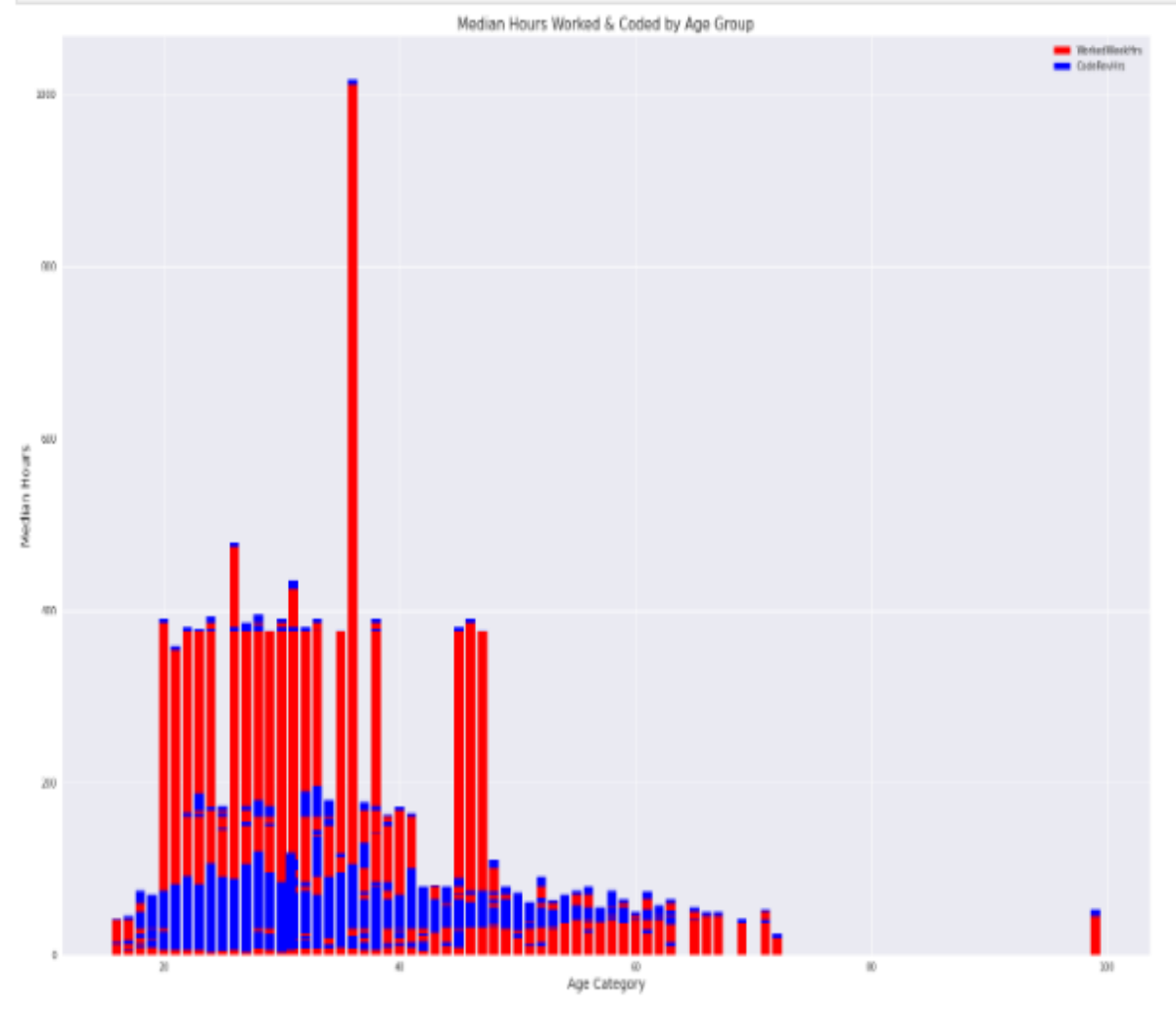
APPENDIX – Developer Analysis



Post- normalization

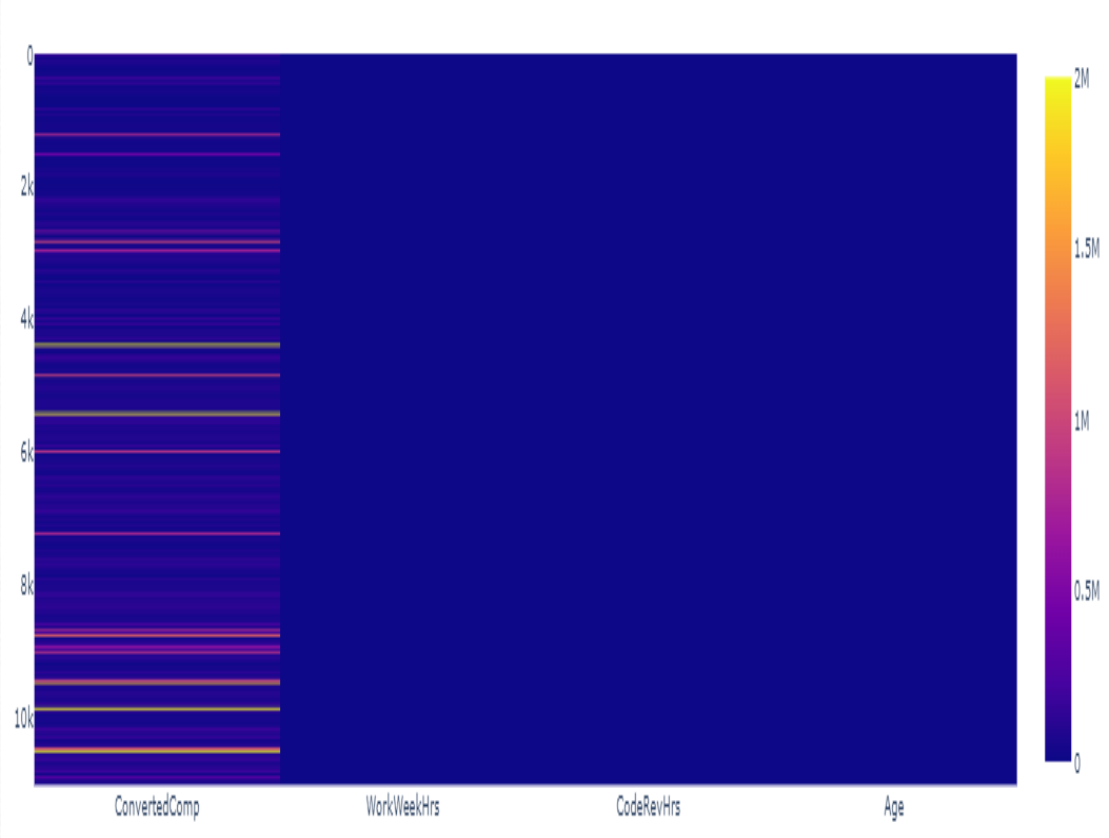
1. Upon normalizing dataset, note that part-time developers are basically non-existent in comparison

APPENDIX – Impact of Multiple Variables



1. Younger techs (<50 years) more dedicated to longer work hours
2. In comparison, younger techs also favour more code review hours

APPENDIX: Relational Analysis-Heat Map



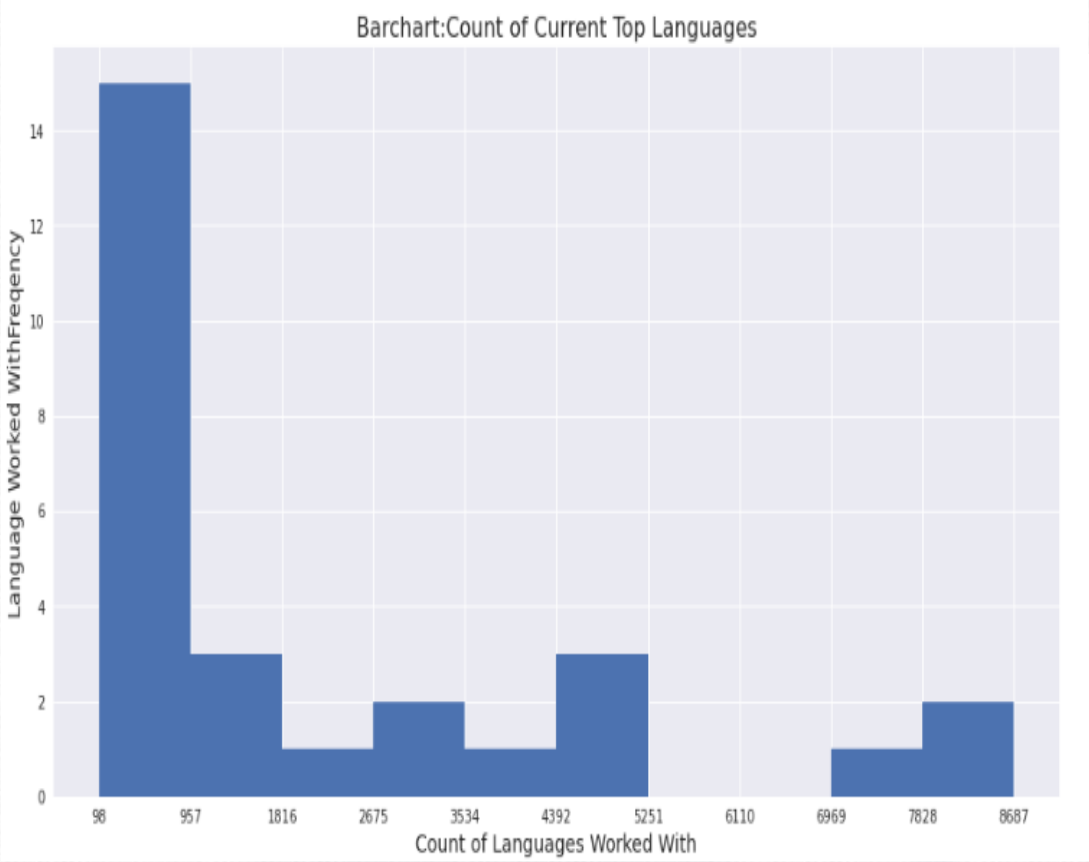
1. Even with attempt to create heat map of the 4 numerical variables, a clear picture not obtained on their correlations

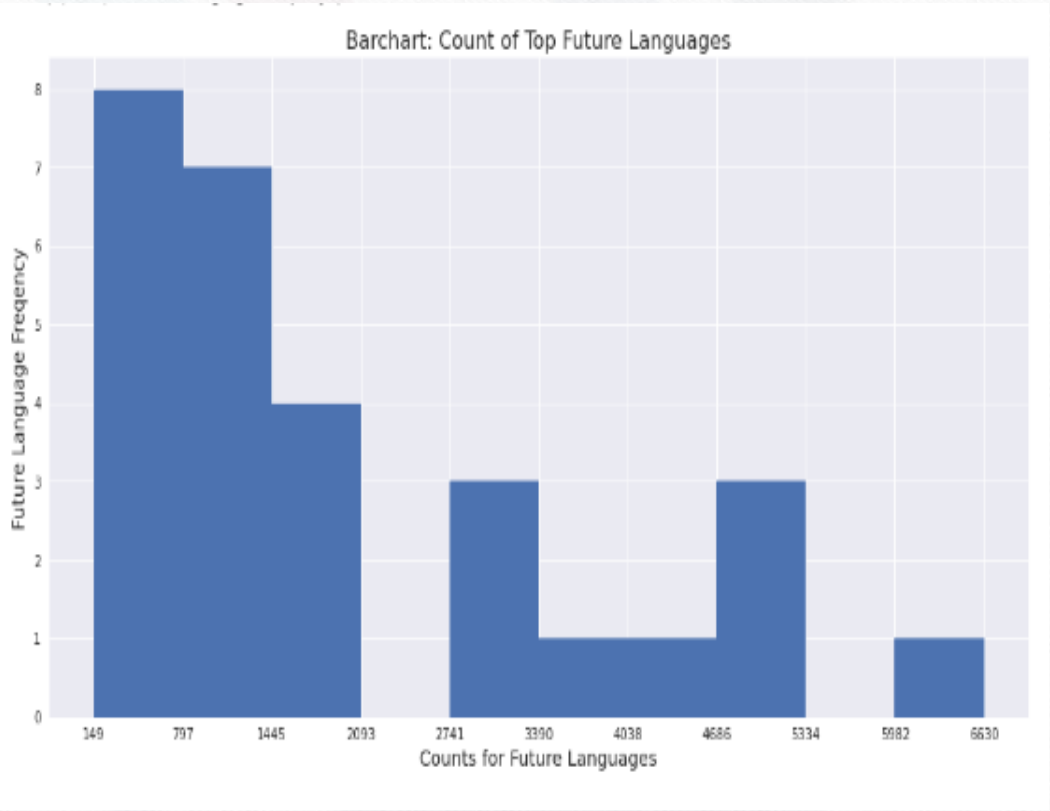
	DevType	cnt
0	Developer, full-stack	6928
1	Developer, back-end	6290
2	Developer, front-end	3920
3	Developer, desktop or enterprise applications	2575
4	Developer, mobile	1959
5	DevOps specialist	1639
6	Database administrator	1413
7	System administrator	1202
8	Designer	988
9	Developer, QA or test	911
10	Developer, embedded applications or devices	854
11	Engineer, data	832
12	Data scientist or machine learning specialist	803
13	Data or business analyst	802
14	Student	766
15	Academic researcher	556
16	Educator	514
17	Product manager	480
18	Developer, game or graphics	472
19	Engineer, site reliability	449
20	Engineering manager	386
21	Scientist	354
22	Senior executive/VP	160
23	Marketing or sales professional	61

APPENDIX: Categorizing IT Tasks

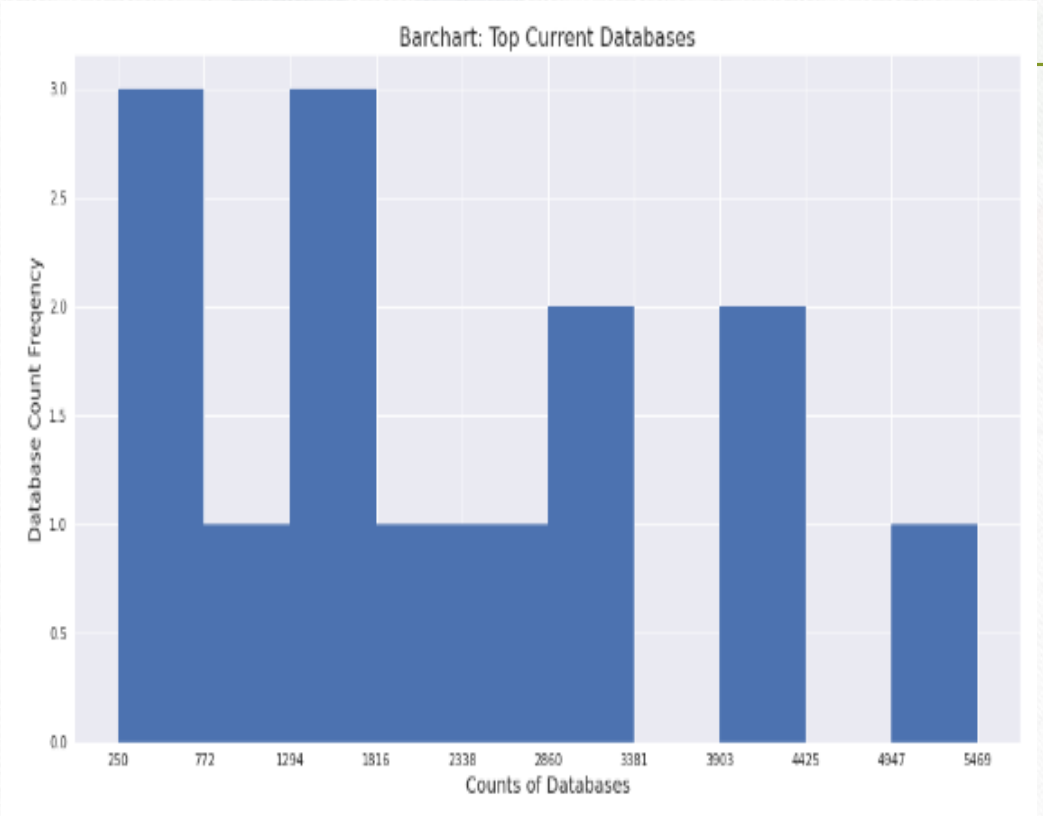
1. Notably full-stack and back-end developers are in high demand
2. Much concern can be expressed for IT marketing/sales arenas with a count of 61

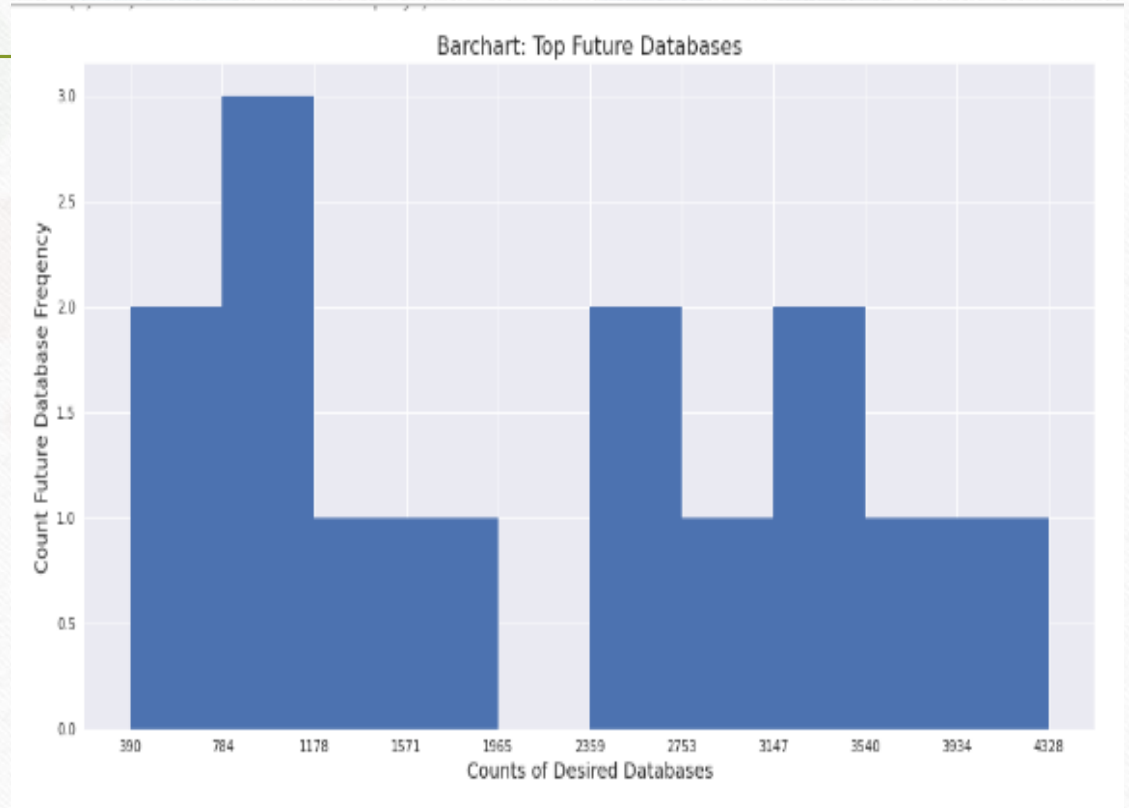
APPENDIX: All Current Language Trends





APPENDIX: All Database Trends

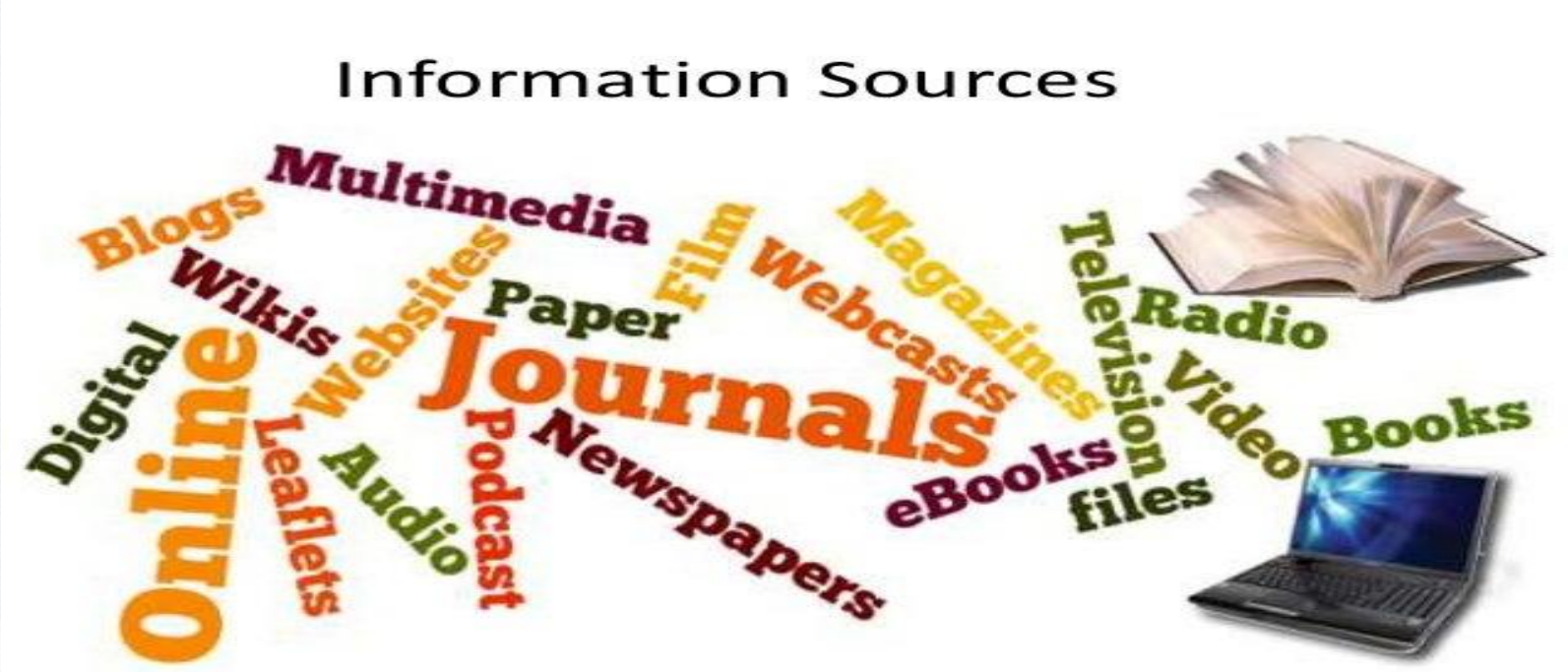




APPENDIX – Video: How to Get a Job in

APPENDIX: Data Source

RESULTS: Data Collection



APPENDIX – Links to IT Datasets

1. https://archive.ics.uci.edu/ml/datasets/KDC-4007+dataset+Collection
2. https://data.world/datasets/information-technology
3. https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DA0321EN-SkillsNetwork/labs/module%201/Exploring%20the%20dataset/About_the_data_set.md.html?origin=www.coursera.org

APPENDIX – Links to Project Codes

1. Data Collection: https://github.com/bahamalytixs/Capstone-Module/blob/main/HR_Collecting_Jobs_data_Using_API-Questions-Final%20(2).ipynb
2. Data Wrangling: https://github.com/bahamalytixs/Capstone-Module/blob/main/HR-M2DataWrangling-Final.ipynb
3. Exploratory Analysis: https://github.com/bahamalytixs/Capstone-Module/blob/main/HR-M3ExploratoryDataAnalysis-lab-Final.ipynb
4. Data Visualization: https://github.com/bahamalytixs/Capstone-Module/blob/main/HR-M4DataVisualization-lab-Final.ipynb

APPENDIX – Data Dictionary 1/6

Column Name	Question Text
Respondent	Randomized respondent ID number (not in order of survey response time)
MainBranch	Which of the following options best describes you today? Here, by "developer" we mean "someone who writes code."
Hobbyist	Do you code as a hobby?
OpenSourcer	How often do you contribute to open source?
OpenSource	How do you feel about the quality of open source software (OSS)?
Employment	Which of the following best describes your current employment status?
Country	In which country do you currently reside?
Student	Are you currently enrolled in a formal, degree-granting college or university program?
EdLevel	Which of the following best describes the highest level of formal education that you've completed?
UndergradMajor	What was your main or most important field of study?
EduOther	Which of the following types of non-degree education have you used or participated in? Please select all that apply.
OrgSize	Approximately how many people are employed by the company or organization you work for?
DevType	Which of the following describe you? Please select all that apply.
YearsCode	Including any education, how many years have you been coding?
Age1stCode	At what age did you write your first line of code or program? (E.g., webpage, Hello World, Scratch project)
YearsCodePro	How many years have you coded professionally (as a part of your work)?
CareerSat	Overall, how satisfied are you with your career thus far?
JobSat	How satisfied are you with your current job? (If you work multiple jobs, answer for the one you spend the most hours on.)
MgrIdiot	How confident are you that your manager knows what they're doing?
MgrMoney	Do you believe that you need to be a manager to make more money?

APPENDIX – Data Dictionary – 2/6

MgrWant	Do you want to become a manager yourself in the future?
JobSeek	Which of the following best describes your current job-seeking status?
LastHireDate	When was the last time that you took a job with a new employer?
LastInt	In your most recent successful job interview (resulting in a job offer), you were asked to... (check all that apply)
FizzBuzz	Have you ever been asked to solve FizzBuzz in an interview?
JobFactors	Imagine that you are deciding between two job offers with the same compensation, benefits, and location. Of the following factors, which 3 are MOST important to you?
ResumeUpdate	Think back to the last time you updated your resumé, CV, or an online profile on a job site. What is the PRIMARY reason that you did so?
CurrencySymbol	Which currency do you use day-to-day? If your answer is complicated, please pick the one you're most comfortable estimating in.
CurrencyDesc	Which currency do you use day-to-day? If your answer is complicated, please pick the one you're most comfortable estimating in.
CompTotal	What is your current total compensation (salary, bonuses, and perks, before taxes and deductions), in CurrencySymbol? Please enter a whole number in the box below, without any punctuation. If you are paid hourly, please estimate an equivalent weekly, monthly, or yearly salary. If you prefer not to answer, please leave the box empty.
CompFreq	Is that compensation weekly, monthly, or yearly?
ConvertedComp	Salary converted to annual USD salaries using the exchange rate on 2019-02-01, assuming 12 working months and 50 working weeks.
WorkWeekHrs	On average, how many hours per week do you work?
WorkPlan	How structured or planned is your work?
WorkChallenge	Of these options, what are your greatest challenges to productivity as a developer? Select up to 3:
WorkRemote	How often do you work remotely?
WorkLoc	Where would you prefer to work?

APPENDIX – Data Dictionary – 3/3

ImpSyn	For the specific work you do, and the years of experience you have, how do you rate your own level of competence?
CodeRev	Do you review code as part of your work?
CodeRevHrs	On average, how many hours per week do you spend on code review?
UnitTests	Does your company regularly employ unit tests in the development of their products?
PurchaseHow	How does your company make decisions about purchasing new technology (cloud, AI, IoT, databases)?
PurchaseWhat	What level of influence do you, personally, have over new technology purchases at your organization?
LanguageWorkedWith	Which of the following programming, scripting, and markup languages have you done extensive development work in over the past year, and which do you want to work in over the next year? (If you both worked with the language and want to continue to do so, please check both boxes in that row.)
LanguageDesireNextYear	Which of the following programming, scripting, and markup languages have you done extensive development work in over the past year, and which do you want to work in over the next year? (If you both worked with the language and want to continue to do so, please check both boxes in that row.)
DatabaseWorkedWith	Which of the following database environments have you done extensive development work in over the past year, and which do you want to work in over the next year? (If you both worked with the database and want to continue to do so, please check both boxes in that row.)
DatabaseDesireNextYear	Which of the following database environments have you done extensive development work in over the past year, and which do you want to work in over the next year? (If you both worked with the database and want to continue to do so, please check both boxes in that row.)
PlatformWorkedWith	Which of the following platforms have you done extensive development work for over the past year? (If you both developed for the platform and want to continue to do so, please check both boxes in that row.)

APPENDIX – Data Dictionary – 4/6

PlatformDesireNextYear	**Which of the following platforms have you done extensive development work for over the past year? (If you both developed for the platform and want to continue to do so, please check both boxes in that row.)**
WebFrameWorkedWith	Which of the following web frameworks have you done extensive development work in over the past year, and which do you want to work in over the next year? (If you both worked with the framework and want to continue to do so, please check both boxes in that row.)
WebFrameDesireNextYear	Which of the following web frameworks have you done extensive development work in over the past year, and which do you want to work in over the next year? (If you both worked with the framework and want to continue to do so, please check both boxes in that row.)
MiscTechWorkedWith	Which of the following other frameworks, libraries, and tools have you done extensive development work in over the past year, and which do you want to work in over the next year? (If you both worked with the technology and want to continue to do so, please check both boxes in that row.)
MiscTechDesireNextYear	Which of the following other frameworks, libraries, and tools have you done extensive development work in over the past year, and which do you want to work in over the next year? (If you both worked with the technology and want to continue to do so, please check both boxes in that row.)
DevEnviron	Which development environment(s) do you use regularly? Please check all that apply.
OpSys	What is the primary operating system in which you work?
Containers	How do you use containers (Docker, Open Container Initiative (OCI), etc.)?
BlockchainOrg	How is your organization thinking about or implementing blockchain technology?
BlockchainIs	Blockchain / cryptocurrency technology is primarily:
BetterLife	Do you think people born today will have a better life than their parents?
ITperson	Are you the "IT support person" for your family?
OffOn	Have you tried turning it off and on again?
SocialMedia	What social media site do you use the most?

APPENDIX – Data Dictionary – 5/6

Extraversion	**Do you prefer online chat or IRL conversations?**
ScreenName	What do you call it?
SOVisit1st	To the best of your memory, when did you first visit Stack Overflow?
SOVisitFreq	How frequently would you say you visit Stack Overflow?
SOVisitTo	I visit Stack Overflow to... (check all that apply)
SOFindAnswer	On average, how many times a week do you find (and use) an answer on Stack Overflow?
SOTimeSaved	Think back to the last time you solved a coding problem using Stack Overflow, as well as the last time you solved a problem using a different resource. Which was faster?
SOHowMuchTime	About how much time did you save? If you're not sure, please use your best estimate.
SOAccount	Do you have a Stack Overflow account?
SOPartFreq	How frequently would you say you participate in Q&A on Stack Overflow? By participate we mean ask, answer, vote for, or comment on questions.
SOJobs	Have you ever used or visited Stack Overflow Jobs?
EntTeams	Have you ever used Stack Overflow for Enterprise or Stack Overflow for Teams?
SOComm	Do you consider yourself a member of the Stack Overflow community?
WelcomeChange	Compared to last year, how welcome do you feel on Stack Overflow?
SONewContent	Would you like to see any of the following on Stack Overflow? Check all that apply.
Age	What is your age (in years)? If you prefer not to answer, you may leave this question blank.
Gender	Which of the following do you currently identify as? Please select all that apply. If you prefer not to answer, you may leave this question blank.
Trans	Do you identify as transgender?
Sexuality	Which of the following do you currently identify as? Please select all that apply. If you prefer not to answer, you may leave this question blank.

APPENDIX – Data Dictionary – 6/6

Ethnicity	**Which of the following do you identify as? Please check all that apply. If you prefer not to answer, you may leave this question blank.**
Dependents	Do you have any dependents (e.g., children, elders, or others) that you care for?
SurveyLength	How do you feel about the length of the survey this year?
SurveyEase	How easy or difficult was this survey to complete?

ACKNOWLEDGEMENTS

1. IBM partners: Providers of course contents
2. Coursera staff: Technical assistance with course materials, production and clarification of data visuals & codes
3. Course participants: for logged questions in forum chats

REFERENCES

1. BrighgtCarbon, "How to Embed a Video in PowerPoint",3 March, 2020, https://www.youtube.com/watch?v=Wq2Lz2kANzU
2. Stackoverflow blog: https://stackoverflow.blog/2019/04/09/the-2019-stack-overflow-developer-survey-results-are-in/ Dataset source
3. Web Photo https://www.shutterstock.com/search/web-development
4. https://www.shutterstock.com/search/data-analysis
5. Remuneration Image - https://www.google.com/search?q=image%3A+money&rlz=1C1HLDY_enBS907BS911&oq=image%3A+money&aqs=chrome..69i57j69i58.7943j0j15&sourceid=chrome&ie=UTF-8
6. Data Dictionary - https://www.google.com/search?q=image%3A+data+dictionary&rlz=1C1HLDY_enBS907BS911&oq=image%3A+data+dictionary&aqs=chrome..69i57j69i58.10349j0j15&sourceid=chrome&ie=UTF-8

REFERENCES

7. Audience Light Applause - https://mixkit.co/free-sound-effects/applause/
8. Data source photo: https://www.google.com/search?rlz=1C1HLDY_enBS907BS911&sxsrf=ALiCzsaOzEl9jykUxaM7bO2xl_n0UQCLrw:1656515047947&source=univ&tbm=isch&q=image:+data+source&fir=JfcKDkObm6D2wM%252CsMNquGiFpxsU5M%252C_%253BHX5BqeDpWDXlCM%252Ccqn6sL6pI14U7M%252C_%253BcyjpzQ14JLGyLM%252CHNUxTGA3ChtGtM%252C_%253BlGD341KtTmF-TM%252CsMNquGiFpxsU5M%252C_%253B9Uh3yWn8KKzC0M%252Cjob0w77G8II1AM%252C_%253BvH6zCax9BpaknM%252C2n_UsbhJvL0LOM%252C_%253BqgI93UJyKGj8xM%252C0eBReWB0beeMCM%252C_%253BqvWrHZtUlNgPZM%252CtHFzUN6qAk0axM%252C_%253BoFr4FFFBiLmAtM%252C5zuDba-4HOnWBM%252C_%253BF9zHOvBWaF1J7M%252CsMNquGiFpxsU5M%252C_&usg=AI4_-kRPprsGtkZlaO0Q4kLMj_TPwLhpEw&sa=X&ved=2ahUKEwixv_2399L4AhUYtoQIHQPfBh0Q7Al6BAgCEEw&biw=1094&bih=504&dpr=1.25

ANIMATED THANK YOU

Thank you

PRESENTATION APPLAUSE