Understanding the Data Science professional – a data driven approach to creating personas of Data Science professionals

Photo by Adeolu Eletu on Unsplash

Introduction

Information Technology (IT) as an industry and its function within an organization has rapidly evolved over the last 25-30 years. The entry of data science, blockchain and AI has rapidly transformed not only the Technology function within the organization but also transformed business, strategy and decision-making processes within the organization, thus spurring technology-led innovations. Roles and job functions within an organization have and are also rapidly evolving. The last Stackoverflow survey (2020) listed eight different developer roles, out of 23 different roles in the survey. We also have mention of specialist roles such as DevOps Specialist, Data Science/Machine Learning Specialist, Data Engineer

and Site Reliability Engineer. The questions that I had after looking at the roles listed in the survey were

1. What are the distinct personas of professionals surveyed by Stackoverflow, given the diversity of the roles listed and the corresponding goals and objectives both for the organization and the professionals?
2. How do these personas relate to compensation levels, both, practically and/or statistically?
3. How do these personas relate to jobsat, both, practically and/or statistically?

1. What are the distinct personas of professionals surveyed by Stackoverflow, given the diversity of the roles and the corresponding goals and objectives both for the organization and the professionals?

The concept of personas has originated from marketing to understand the customer. In recent times, this concept has gained wider acceptance across domains, spanning both academic and practitioner communities as an interactive product design technique.

As per Goodwin, personas are fictional, detailed, archetypal characters that represent distinct groups of behaviors, goals and motivations observed and identified during the research phase [1]. This effectively means that a persona is

- a fictitious user that has the characters of a group of similar (real) users.
- defined by his/her goals or objectives or area(s) of focus
- created by analyzing goals, objectives, areas of focus/ interest, motivations and behaviors of real users. This data may be collected from marketing research, user studies including interviews and questionnaires.

The persona creation process is mostly viewed as a process driven by qualitative analysis of market research data. However, quantitative methods such as cluster analysis are increasingly being adopted in the persona creation process.

The dataset used for our analysis in this blogpost is from Stackoverflow's 2020 Annual Survey. The survey data contains survey responses from 64,000 respondents across 184 countries. The survey aimed to understand multiple aspects of jobs related to entire spectrum of digital solution development lifecycle, across practitioner and academia, and across functions including Sales and Marketing, Research, Engineering, Technology Operations and across multiple levels – from developers to middle and senior Management.

Given the diversity and maturity levels of business and technology functions across countries and regions, I decided to limit my analysis to a single country – United States. With 12,000+ respondents from US alone, this also provided a good volume of data for the analysis exercise.

The questions in the survey that are particularly relevant to us considering the questions that I have listed above are:

a) Which of the following describe you? Please select all that apply.
b) Salary converted to annual USD salaries using the exchange rate on 2020-02-19, assuming 12 working months and 50 working weeks.
c) How satisfied are you with your current job? (If you work multiple jobs, answer for the one you spend the most hours on.)

The survey response for (a) allowed for selection of multiple options for each respondent and stored in a single data column. The survey responses for (b) and (c) allowed for entry/ selection of a single value. Job satisfaction (jobsat) levels were measured on a 5 point scale.

Before we get into the analysis findings, here is some brief information on the approach adopted for this analysis:

- Data Cleansing and Missing value treatment
  - A separate dataset of observations containing survey responses from United States was created as input data for the analysis exercise
  - Only observations where responses for (a) (attribute *DevType*) and (b) (attribute *ConvertedComp*) were retained; remaining observations were deleted from the dataset. Missing values for the converted compensation were not imputed since this number was significant (about 30% of the observations) and imputing it to a mean value would have distorted the input dataset
  - Since the survey response for (a) allowed for selection of multiple options for each respondent and stored in a single data column, a list containing unique set of roles selected by respondents was extracted from the dataset; this resulted in a list of 23 unique roles.
- Feature Engineering
  - A categorical variable was created for each of the unique roles in the above-mentioned list; each of these categorical variables were set to 1 if role was selected by the respondent and 0 otherwise
  - Since we need numerical features for clustering, these were created by target encoding the values for compensation level (attribute *ConvertedComp*). These set of target encoded variables for each of the 23 unique roles (engineered features) were used as inputs for clustering
- Pre-processing
  - The engineered feature dataset was then scaled; RobustScaler was used since this scaler handles outliers well compared to other scalers
  - Before clustering the data, Principal Component Analysis (PCA) was performed on the scaled dataset to reduce the number of dimensions. The number of dimensions required to explain at least 80% of the variance was determined; 8 dimensions were adequate to explain 82% of the variance.
- Clustering

- K means clustering was then performed on the reduced data that was obtained from the PCA. The K means clustering algorithm was run on three alternative reduced datasets (datasets containing 7, 8 and 9 dimensions). Silhouette score was used as the metric to compare various clustering models. Since we could achieve a better silhouette score with 7 dimensions, this dimensionally reduced set of 7 features were used as inputs for the clustering algorithm.
- The dataset containing 7 dimensions was used for conducting clustering trials – using different number of clusters as input , ranging from 2 to 12 clusters; intent being to find the optimal number of clusters using the elbow method. Based on this method, with a silhouette score of 0.79, an optimal clustering result could be achieved with 8 clusters.
- The 8 cluster model was then saved, and used to predict the cluster numbers for each observation in the dataset

- Analyze results and document findings
  - The predicted cluster groups and attributes relevant for our analysis (compensation levels and jobsat levels) were then aggregated and visualized to seek answers to the questions listed in this blogpost
  - Relevant statistical tests were conducted to assess the statistical significance of conclusions drawn from the aggregated and visualized data

A diagrammatic representation of the above-mentioned approach is shown in Exhibit 1 below.

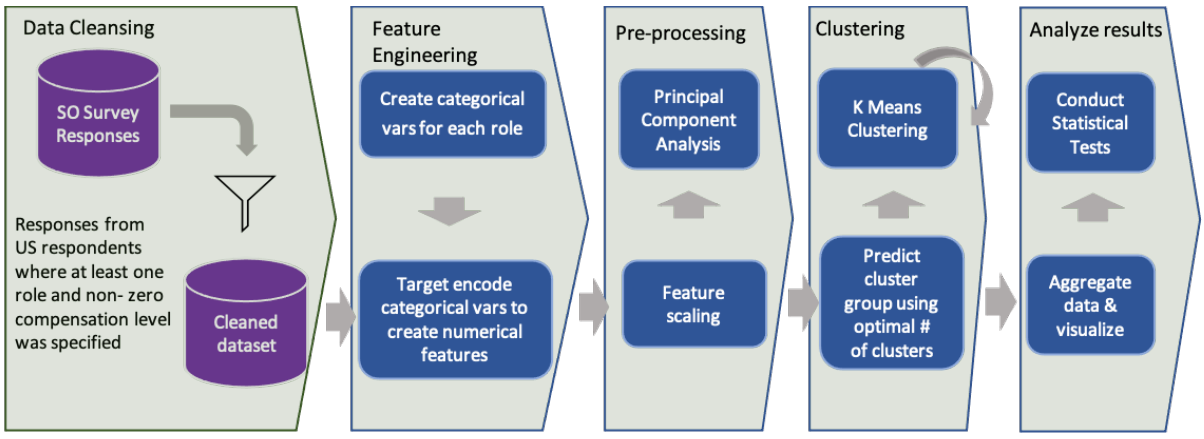


Exhibit 1 – Quantitative Analysis Approach

A heatmap of the degree of presence of the 23 roles in the Stackoverflow survey across various cluster groups is shown in Exhibit 2 below:

Exhibit 2 - Role Types across groups

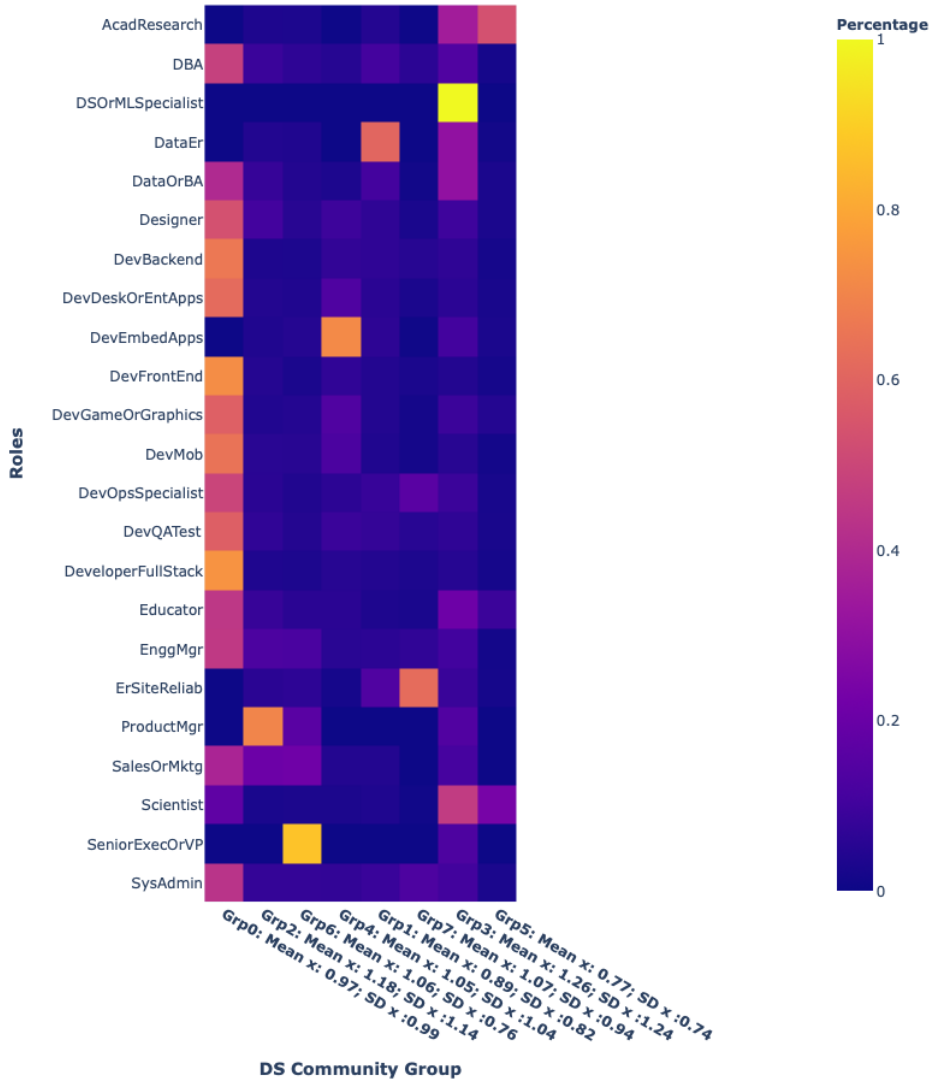


Exhibit 2 provides us with insights about both, roles that are represented and those that are not represented in each group/ cluster resulting from our cluster analysis. Based on this information, we can draw conclusions regarding the nature of activities undertaken by the professionals in the group and based on the same draw conclusions regarding the group persona. Key observations and takeaways are tabulated below:

Group	A - Roles not represented	B - Roles represented	Nature of activities undertaken	Group Persona
0	▪ Academic Researchers, DS or ML Specialist, Data Engineer,	▪ All other roles except those listed in column (A) - 40-60%	This group comprises of professionals that undertake activities spanning the entire spectrum of front-to-	Front-to-back, multi-channel, digital solution development professionals

Group	A - Roles not represented	B - Roles represented	Nature of activities undertaken	Group Persona
	Developer - Embedded Apps, Site Reliability Engineers, Product Manager, Senior Exec or VP		back, multi-channel, digital solution development lifecycle – conceptualization through to sales and marketing	
1	▪ DS or ML Specialist, Product Manager, Senior Exec or VP	▪ Data Engineer (61%) ▪ Site Reliability Engineer, Sys Admin, Data Analyst/ BA, DBA (9—14%) ▪ Scant presence all other roles - <5%	The large presence of Data Engineers in this group indicates that professionals in this group are engaged in developing production grade data pipelines, production site reliability assessment, system administration and maintenance activities.	Data product/ solutions deployment and maintenance professionals
2	▪ DS or ML Specialist, Senior Exec or VP	▪ Product Manager (70%) ▪ Sales and Marketing (20%) ▪ Engineering Manager (13%) ▪ All other roles have scant presence (<=10%)	Significant presence of product managers, besides presence of sales and marketing and engineering managers indicates professionals in this group are engaged in tactical management activities such as translating ideas into marketable products, serving as a bridge between technologists and senior management, and across functional and/or organizational boundaries	Middle management professionals across Engineering, Product Management, Sales & marketing functions
3	▪ None	▪ DS or ML Specialist (100%) ▪ Scientist, Academic researcher,	With all the professionals engaged in DS or ML Specialist roles belonging to this group, besides large presence of Data Engineers, Data	Professionals engaged in conceptualization, development, deployment, maintenance and

Group	A - Roles not represented	B - Roles represented	Nature of activities undertaken	Group Persona
		Data Engineer, Data/ Business Analysts – 30-45% ▪ General and Senior Management roles – Educator, Product Manager, Engineering Manager, Senior Executive/VP, Sales & Marketing – 10-20%, ▪ All other roles - < 10%	and Business Analysts, and product engineering, management and sales, indicates that this group comprises of professionals that undertake activities spanning the entire spectrum of data i.e. AI/ML products/ solution development lifecycle	sales/ marketing of data products/ solutions
4	▪ DS or ML Specialist, Senior Exec or VP, Product Manager	▪ Developer - Embedded apps - 71%, ▪ Developers - Desktop or Enterprise Apps , Game or Graphics, Mobile developers - ~14%	Significant presence of embedded application developers indicates professionals in this group are primarily engaged in development of embedded applications/ solutions	Professionals engaged in development of embedded application solutions
5	▪ DS or ML Specialist, Senior Exec or VP, Product Manager	▪ Academic research - 54% ▪ Scientist - 24% ▪ All other roles – 1-3%	Significant presence of academic researchers and scientists in this group indicates professionals in this group are primarily engaged in data science/ AI related R & D activities	Professionals in academic institutions, data science R & D
6	▪ DS or ML Specialist	▪ Senior Executive/ VP – 87%	Very significant presence of Senior management professionals in this group, besides presence	Senior management professionals/ Executive level

Group	A - Roles not represented	B - Roles represented	Nature of activities undertaken	Group Persona
		▪ Sales and Marketing – 22% ▪ Product Manager, Engineering Manager – 12%-16% ▪ All other roles – 1-3%	of managers across functions such as Product Management, Engineering and Sales and Marketing indicates that professionals in this group are engaged in strategic management activities.	professionals across Engineering, Product Management, Sales & marketing functions
7	▪ DS or ML Specialist, Data Engineer, Senior Exec or VP, Product Manager, Sales & Marketing, Scientist, Academic Researcher, Embedded Apps developer	▪ Site Reliability Engineer – 62% ▪ DevOps Specialist, System Administration- 13-16% ▪ Engineering Manager, DBA – 5-7% ▪ All other roles – <5%	Significant presence of site reliability engineers, besides presence of DevOps specialists, System Administrators and DBAs indicates that professionals in this group are engaged in management of technology operations.	Technology Operations management professionals

2. How do these personas relate to compensation levels, both, practically and/or statistically?

Let us try to seek answers to this question from macro and micro perspectives. These are presented in Exhibits 3 and 4.

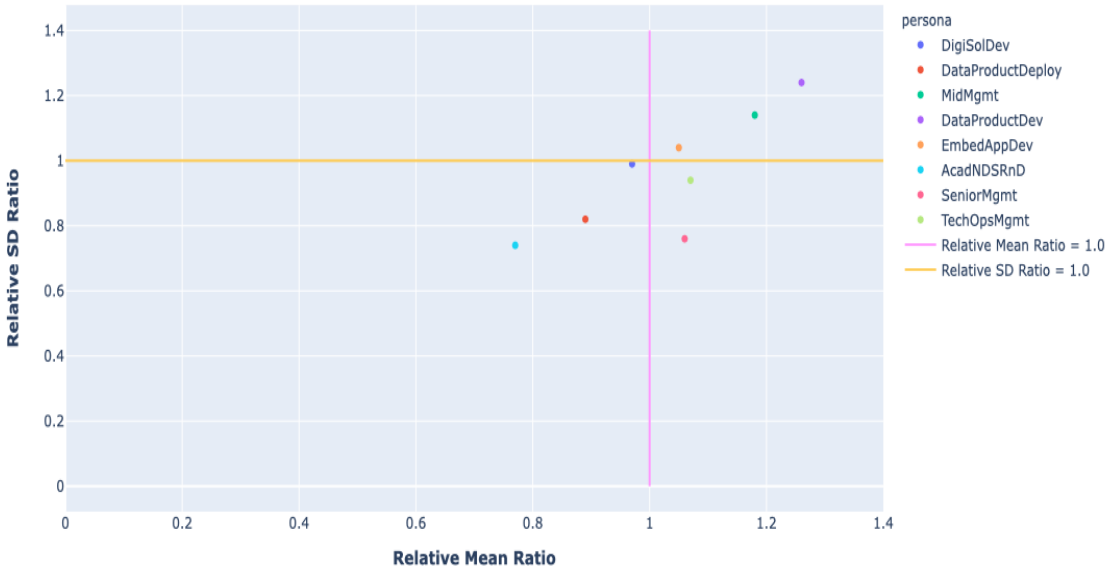
First the macro perspective; let us look at the relative mean and variability (standard deviation) of the compensation levels in each group by comparing the group mean and standard deviation with the mean and standard deviation at the country/national level. Based on this criteria we have four quadrants. Let us look at the groups that fall in each of these quadrants:

- Low relative mean, low relative variability : Three groups - Professionals in academic institutions, data science R & D, Front-to-back, multi-channel, digital solution development professionals and, Data product/ solutions deployment and maintenance professionals belong to this quadrant
- High relative mean, high relative variability: : Three groups - Middle management professionals across business functions such as Engineering, Product Management, Sales & Marketing, professionals engaged in conceptualization, development, deployment, maintenance and sales/ marketing of data products/

solutions and, professionals engaged in development of embedded application solutions belong to this quadrant

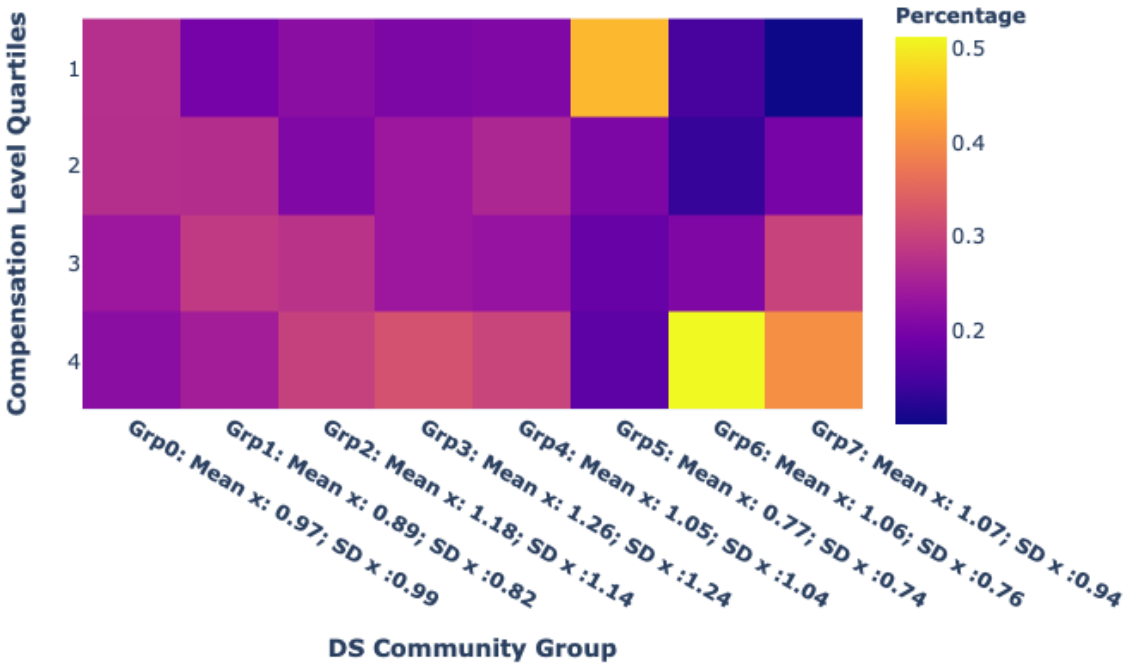
- High relative mean, low relative variability: Two groups - Senior management professionals/ Executive level professionals across functions such as Engineering, Product Management, Sales & Marketing and, Technology Operations Management professionals belong to this quadrant

Exhibit 3 – Macro perspective: Relative comparison of compensation levels across groups



Let us now take the micro perspective and look at the percentage membership of each group in each quartile of compensation level; the quartiles being calculated by ranking the compensation levels at the country/national level i.e. across groups. This is presented in Exhibit 4.

Exhibit 4 - Compensation Levels across groups



Key takeaways

- Groups 0 (front – back solution development professionals) and 5 (academic and data science R & D professionals) have compensation mean and standard deviation ratios below 1 i.e. group mean and standard deviation levels are below country mean and standard deviation respectively. In these groups more than 50% of the group members have compensation levels below 50 percentile
- Group 1 (data science solution deployment and maintenance professionals) is a special case – here group compensation mean and standard deviation ratios are below 1 but 54% of the members in the group have compensation levels above 50 percentile.
- Groups 6 and 7 have group compensation mean levels above 1 but standard deviation ratio is below 1, indicating higher relative compensation levels and lower relative variation. In these groups more than 70% of the members have compensation levels above 50 percentile.
- Groups 2, 3 and 4 have group compensation mean and standard deviation ratios above 1, indicating higher relative compensation levels and greater relative variation. 53% - 56% of the members in these groups have compensation levels above 50 percentile; this is close to the levels seen for Group 1 where group compensation mean and standard deviation ratios are below 1 yet 54% of the members in the group have compensation levels above 50 percentile.

With a better practical understanding of the relationship from practical perspective, we also assess if the relationship between compensation levels (expressed in the form

of a categorical variable) and another categorical variable (cluster groups/ personas) are statistically significant. This can be done using a Chi squared test of independence. We find that they are indeed related – with the null hypothesis (convincingly) rejected.

3. How do these personas relate to jobsat, both, practically and/or statistically?

We look at the relative job satisfaction levels in each group by comparing the percentage job satisfaction level (within the group) with the corresponding job satisfaction percentage at the country level. This is presented in Exhibit 5 below:

Key takeaways:

- Professionals in groups 6, 5 and 3 outperform corresponding national job satisfaction levels – more professionals are Very Satisfied and less professionals

are Dissatisfied as compared to corresponding national percentages. The percentages of professionals who are Very Dissatisfied is also close to national percentage ~+/- 1-1.5% than national percentage.

- The job satisfaction levels of professionals in Group 0 are pretty much in line with national levels for all levels of job satisfaction (and dissatisfaction).
- In terms of being Very Satisfied, the relative percentages for professionals in the remaining groups i.e. groups 7,4, 2 and 1 are below national percentages. However, the professionals in groups 1 and 7 outperform national figures in terms of being 'Slightly Satisfied' with their job and fewer professionals are dissatisfied with their jobs than corresponding national percentage figures, indicating a positive overtone to job satisfaction levels as compared to national percentage.
- Professionals in groups 2 and 4, though have higher dissatisfaction levels than corresponding national percentage figures, indicating a not so positive overtone to job satisfaction levels as compared to national percentage.

In summary, in terms of increasing levels of relative jobsat levels as compared to corresponding national percentages, we can conclude that professionals in groups 2 & 4 are not so positive about job satisfaction levels, professionals in group 0 are pretty much in line with national levels for all levels of job satisfaction (and dissatisfaction), professionals in groups 1 and 7 are *Slightly Satisfied* and professionals in groups 6,5 and 3 are the happiest lot with higher relative jobsat levels compared to corresponding national levels.

Having said that, the next logical and related question that does crop up is that, from a statistical perspective, are jobsat and cluster group variables dependent or independent of each other. A chi-squared test of independence was therefore conducted to seek an answer to this question. The answer is – yes they are dependent (the null hypothesis of them being independent was rejected).

Conclusion

In this post, we took a quantitative approach to creation of personas of Data Science professionals using robust dataset from Stackoverflow's 2020 survey. We conclude that:

a) The 23 roles listed by the respondents could be grouped into 8 distinct personas –
   i) Front-to-back, multi-channel, digital solution development professionals
   ii) Data product/ solutions deployment and maintenance professionals
   iii) Middle management professionals across Engineering, Product Management, Sales & marketing functions
   iv) Professionals engaged in conceptualization, development, deployment, maintenance and sales/ marketing of data products/ solutions
   v) Professionals engaged in development of embedded application solutions
   vi) Professionals in academic institutions, data science R & D,
   vii) Senior management professionals/ Executive level professionals across Engineering, Product Management, Sales & marketing functions,
   viii) Technology Operations management professionals

b) The above personas are related to compensation levels practically and the relationship is also statistically significant.
c) We also explored the relationship between these groups/ personas and job satisfaction levels and found that they are practically related and the relationship is also statistically significant

References

1. Cheat sheet for implementing 7 methods for selecting the optimal number of clusters in Python https://towardsdatascience.com/cheat-sheet-to-implementing-7-methods-for-selecting-optimal-number-of-clusters-in-python-898241e1d6ad
2. One Hot Encoding, Standardization, PCA: Data preparation for segmentation in python https://towardsdatascience.com/one-hot-encoding-standardization-pca-data-preparation-steps-for-segmentation-in-python-24d07671cf0b
3. A refresher on statistical significance https://hbr.org/2016/02/a-refresher-on-statistical-significance
4. A gentle introduction to the Chi-squared test for Machine Learning https://machinelearningmastery.com/chi-squared-test-for-machine-learning/