**DATA PREPARATION PLAN – STACK OVERFLOW**

**Topic:** Stack Overflow Data Analysis

**Data Source:** Kaggle

**Data Source Link:** [https://www.kaggle.com/](mailto:https://www.kaggle.com/)

**Data URL** : <https://www.kaggle.com/stackoverflow/stack-overflow-2018-developer-survey>

**1.**

**About Data Source:**

Kaggle is an online platform to compete with others in competitions which are based on data sources, datasets, data manipulation over machine learning tasks. The key fact is the Kaggle is open to multiple utilities like competitions, data challenges, data sourcing, data providing and many more. The competition are only related to machine learning, data science, Deep learning or AI related. Most of time you have been given some training and testing dataset to build some good machine learning models and when you public your kernel , others will review it and then gives you upvotes. The data owners who agree to the terms of the platform globally come across at the platform and share their data for an open source usage for users worldwide for the good.

**About the data :**

This dataset contains the survey results of approx. 20000 developers having information about their favorite technologies to their job preferences. This dataset has 129 columns and 98,855 rows of qualitative data about the developers.

The data contains insights such as “Countries to which developers belong to”, “Salaries developers were receiving”, “years of work experience”, ”Coding languages interested to work with”,” If developers are looking for a job change”, ”Areas of expertise”, ”Job satisfaction level” and many more.

This 2018 Developer Survey results are organized on Kaggle in the following 2 tables:

1. **survey\_results\_public** contains the main survey results.
2. **survey\_results\_schema** contains each column name from the main results along with the question text corresponding to that column(METADATA).

**License or terms of use:** <https://opendatacommons.org/licenses/odbl/1.0/>

**ODC Open Database License (ODBL) -** <https://opendatacommons.org/licenses/odbl/1.0/>

**2.**

**Intellectual policy constraints, or lack through the Kaggle terms**

License, Policy and Terms of Use of Kaggle are quite extensive and comprehensive and mostly covers all aspects which makes it difficult to find loopholes, constraints or lacks. However, we were able to list a few findings before which we would like to state a glimpse of terms, policy and preamble.

**Terms of Use**

Terms of Use (the “Terms”) are a binding contract between you and Kaggle Inc. (“Kaggle,” “we” and “us”). You must agree to and accept all of the Terms, or you don’t have the right to use the Services. Your use of the Services in any way means that you agree to all of these Terms, and these Terms will remain in effect while you use the Services. These Terms include the provisions in this document, as well as those in the Privacy Policy and Copyright Dispute Policy.

You represent and warrant that you are of legal age to form a binding contract. If you are not of legal age to form a binding contract, please see the Privacy Policy and contact us at Kaggle Support for more information on the parental consent process. You must either be of legal age to form a binding contract or have completed the parental consent process to use the Services.

If you’re agreeing to these Terms on behalf of an organization or entity, you represent and warrant that you are authorized to agree to these Terms on that organization or entity’s behalf and bind them to these Terms (in which case, the references to “you” and “your” in these Terms, except for in this sentence, refer to that organization or entity), but you are still required to sign up each individual user from your organization with a Kaggle User ID. You are not allowed to allow multiple individuals to operate under one Kaggle User ID, even if you are an organization or entity.

**Privacy Policy**

This Privacy Policy is meant to help you understand what information we collect, why we collect it, and how you can update, manage, export, and delete your information. If you have any questions about this Privacy Policy, you can contact us here.

When you clicked “accept” or “agree” in connection with registering your account, we made available to you this Privacy Policy and the information handling practices described in here. This Privacy Policy only covers Kaggle's practices regarding personal information, and not information more generally. We may use aggregated, anonymized data that we derived from your personal information before you deleted it, but not in a manner that incorporates any of your personal information or would identify you personally.

We’ll ask for your consent before using your information for a purpose that isn’t covered in this Privacy Policy. You may, of course, decline to share certain personal information with Kaggle, in which case Kaggle may not be able to provide to you some of the features and functionality found on the Services.\\

**Preamble**

The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Database while maintaining this same freedom for others. Many databases are covered by copyright, and therefore this document licenses these rights. Some jurisdictions, mainly in the European Union, have specific rights that cover databases, and so the ODbL addresses these rights, too. Finally, the ODbL is also an agreement in contract for users of this Database to act in certain ways in return for accessing this Database.

Databases can contain a wide variety of types of content (images, audiovisual material, and sounds all in the same database, for example), and so the ODbL only governs the rights over the Database, and not the contents of the Database individually. Licensors should use the ODbL together with another license for the contents, if the contents have a single set of rights that uniformly covers all of the contents. If the contents have multiple sets of different rights, Licensors should describe what rights govern what contents together in the individual record or in some other way that clarifies what rights apply.

Below mentioned are a list of Intellectual policy constraints, or lack through the Kaggle terms:

1. The terms of use determine conditions when you sign up to Kaggle for use however it is not explicitly mentioned about third party platforms intervening through the user usage to extract data and make use of the data.
2. It states that “You represent and warrant that you are of legal age to form a binding contract” however no legal or minimum age is referenced here for the usage purpose.
3. Multiple users are not allowed to use through a single sign in, however there is no multifactor authorization which prohibits multiple users to access through different locations or IP addresses.
4. The privacy policy confines all actions to be performed by the user however, does not restricts the user for further trading of data which should be unethical and illegal as per a data point of view.

**3.**

**Description of the Metadata:**

For our data set a file with the description of the variables was available through Kaggle. The documentation shows the variable name and what the question that each individual was asked. For the most part, most if the variables that we are using are self-explanatory, but if there were some confusion, we can always use the metadata to figure out what question was asked for each column.

In addition to the question, some variables have additional documentation on what type of answers can be expected from users. For example, some questions are “select all that apply”, some are open text answers, and some have the option to be left blank. This gives us additional information on why users interviewed answered the way they did. Overall, the documentation on each variable that we plan on using, were very helpful in us understanding the data that we are working with.



**4**

**Issues Encountered with data:**

1. There were a lot of unidentified characters and symbols like â€™. These are the data chunks which are having very indifferent ASCII values than the alphanumeric characters as also they provide no use while data analysis.
2. The unidentified characters and symbols act as a roadblock and hindrance for data cleaning, data delimiting and data formatting since they intervene and result to null data points in the fields.
3. There are a lot of NULL data observed through the dataset. The NULL data plays a crucial role in a dataset in order to identify values holding no values or zero values or junk values. The null values are hard to play with since they have to be kept at some places to determine the count of observation and at the same time need to be removed or eliminated in order to run statically analysis.
4. Open ended text data are meant for observational analysis since no statistics or data analytics can be performed over them. Moreover, these kinds of input vary from each observation and variability is huge resulting in nil or least correlations, hence they cannot be grouped or analyzed.
5. The open-ended text need to be ignored at places where mathematics or grouping is required, as also it can be extremely useful if we run a text analysis or sentiment analysis to determine the nature of input.

**5.**

**Remediation to issues encountered**

There are various ways how each issue which is evident in the beginning and which rises up over the time through analysis is analyzed, interpreted and worked upon. Below are the actions taken by us for remediation of the issues encountered:

1. The unidentified characters through the data are delimited through the recognized symbols and are split into cells and then using R functions, those rows with the symbols are eliminated or kept separately in other column, unperturbed, for any further utility if required. “Symbols might always have a meaning”.
2. The NULL data is replaced by values of zeros where count was required along with mathematical importance. However, using R functions, they can be ignored while performing any statistical analysis while they do not disturb our data frame or results.
3. Open ended texts can be recoded to be given likert values for grouping and better understanding over their essence which would not only help us understand their meaning better but also run some analysis and work upon them.
4. NULL values from the primary key are removed since they hold no importance and shall not be included into the analysis.
5. There were a few observations which were misplaced through the matrix which were fixed by using indexing values to the attributed and identifying points of deviations or conflicts.

**6.**

**Data Cleaning Process**

The data consists of 129 columns and 98,855 rows. Out of these 129 columns only selective columns are needed for analysis for that extensive data cleaning was performed to cater following research questions:

**Research Questions**

1. Job satisfaction comes with more salary compensations/expectations?
2. If large Multinational Corporations want to offshore their workload, which countries can be suitable in terms of logistics and workforce?
3. Does location affect the skillsets and technical expertise?
4. Does the number of years in a job affect the workers’ willingness to switch?
5. To check stability in a person based on years of work, competencies, age, salary and last switch to control attrition.
6. Do people use sites like stack overflow to find new jobs?

The process of data cleaning involved following steps:

1. **Identified the columns which were necessary and could be used to provide analytical statistics for the above mentioned research questions.**

We had 6 research questions to cater to, hence we selected only those columns which could help us find the analytical statistics for those questions.

Following are the research questions along with the data columns associated for them:

1. ***Job satisfaction comes with more salary compensations/expectations?***

**Columns Required:**

Country, Employment, Company Size, DevType, Years of Coding Professionally, Job/Career Satisfaction, Hope five years, Job search status, last new job.

1. ***If large Multinational Corporations want to offshore their workload, which countries can be suitable in terms of logistics and workforce?***

**Columns Required:**

Country, Formal education, undergrad major, DevType, years coding professionally, communication tools, education types ,languages worked upon, database/platform/framework worked with, IDE, Operating System

1. ***Does location affect the skillsets and technical expertise?***

**Columns Required:**

Country, Languages Worked With, Database Worked With, Platform Worked With, Framework Worked With

1. ***Does the number of years in a job affect the workers’ willingness to switch?***

**Columns Required:**

Country, Employment, Company Size, DevType, Years of Coding Professionally, Job/Career Satisfaction, Salary, Last New Job, Salary Type, Currency

1. ***To check stability in a person based on years of work experience, competencies, age, salary and last switch to control attrition***.

**Columns Required:**

Country, Employment, Company Size, DevType, Years of Coding Professionally, Job/Career Satisfaction, Salary, Last New Job, Salary Type, Currency, Age, Gender, Hope five years, Job search status

1. ***Do people use sites like stack overflow to find new jobs?***

**Columns Required:**

Country, Employment, Company Size, DevType, Years of Coding Professionally, Job/Career Satisfaction, Salary, Last New Job, Job search status, StackoverflowHasAccount, StackOverflowJobs.

1. **Got rid of the unnecessary columns.**

Initially there were 190 columns and after cleanup we had 54 columns. We did this by indexing the variables that we needed and storing it to a new data frame through R.

1. **Identifying undefined characters in the dataset and replaced them with logical characters.**

In few columns we had undefined characters like â€™. We replaced these undefined characters with logical characters. For example, “Iâ€™m not actively looking, but I am open to new opportunities”. Through R scripts, we replaced the character with “a” to make the sentence “I **am** not actively looking, but I am open to new opportunities”

1. **Catered the null values in the data set by replacing it with NA.**

With the use of R scripts, we replace all NA values in the data frame with the value “NA”

# Works Cited

Kaggle. (2018, NA NA). *Terms of Use*. Retrieved November 7, 2018, from kaggle.com: https://www.kaggle.com/terms

Open Data Commons. (NA, NA NA). *Open Database License (ODbL) v1.0*. Retrieved November 7, 2018, from Open Data Commons: https://opendatacommons.org/licenses/odbl/1.0/

Overflow, S. (2018, May 1). *Stack Overflow 2018 Developer Survey.* Retrieved November 7, 2018, from Stack Overflow: https://www.kaggle.com/stackoverflow/stack-overflow-2018-developer-survey/home

**Word count – 2157**