

REPORT ON daily engagement udacity dataset



By/ Mostafa Noaman Ahmed

|  |  |  |
| --- | --- | --- |
| **S. NO.** | **Topic** | **Page No.** |
| 1 | Introduction   * Business Problems We are trying to solve * Main goal * About Dataset   + Table1- DataSet Preview   + Table 2 - Columns in Dataset | 3  4  4  4  4  4 |
| 2 | EDA & Business Implication   * Analysis of the Data * Table 3- Checking of the Skewness * Univariant Analysis – Box Plot   + Plot 1- Box Plot * Univariant Analysis of Each Column   + Plot 2 – Analysis of acct   + Plot 3- Analysis Of has\_visited   + Plot 4- Analysis of projects\_completed   + Plot 5- Analysis of lessons\_completed   + Plot 6- Analysis of total\_minutes\_visited * Univariant Analysis – Bar Plot   + Plot 7 – Bar Plot * Univariant Analysis of Important relations * Plot 8 – course studying hours * Plot 9 – studied lessons per course * Plot 10 – done projects per course * Plot 11 – most vs least studying users * Plot 12 – lessons studied by users * Plot 13 – projects done by users | 5  5  5  7  7  7  7  7  7  7  7  9  9  9  9  9  10  11  12  13  14 |
| 3 | User progress per each year   * Getting user metadata by code * Regex for cleaning data * Distributed users’ data per year * Univariant Analysis of Important relations for year 2014 – Bar Plot * Plot 13 – Intro to Data Science * Plot 14 – Data Visualization and D3.js * Plot 15 – Data Analysis with R * Plot 16 – Javascript Basics * Plot 17 – Intro to Machine Learning * Plot 18 – A/B Testing * Plot 19 – Data Wrangling with MongoDB * Plot 20 – Intro to HTML and CSS * Plot 21 – Data Analyst Nanodegree | 15  15  16  17  18  18  18  19  19  20  20  21  21  22 |
| 4 | Final Interpretation/Recommendation   * Final Words | 22  22 |

1. **Introduction**

Udacity is a really famous platform for different learning fields especially the technical topics which are related to programming in general, and “Data Science” in particular, we are here trying to analyze the daily engagement dataset to got the knowledge about the most important topics and to know each user interests to meet their needs.

**Business Problems We are Trying to Solve**

* The most nanodegrees that users are interested in.
* The preference of each user per the nanodegrees.
* The different and common users for the same nanodegrees.
* The studying statistics related to each user along the year.

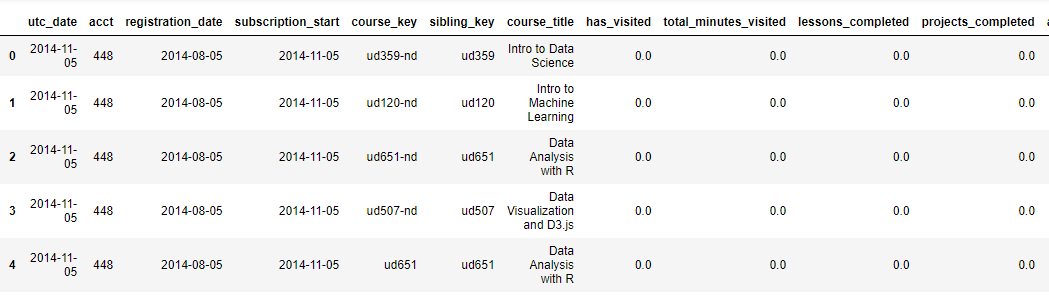
**Main Goal**

* Create an analytical framework to understand
  + User interests
  + Type of Topics that most users prefer

**About Dataset**

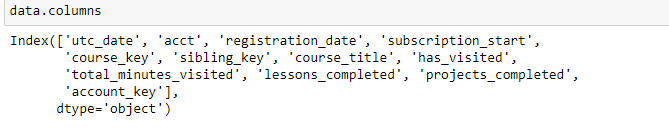
The dataset we’ve consisted of records of users’ daily engagement to different Udacity Nanodegrees. It has various columns which describe the daily engagement such as each user account key, course title, total minutes spent, number of lessons completed, number of projects completed, etc. Such a dataset is very useful for Udacity company and the content creators works in it, on which different analysis can be done to get insights about the users and also the popular courses which attracts users

This is what the dataset looks like:



**Table1- DataSet Preview**

It has lots of rows and columns, which is 2309239 rows spread across 12 columns. Here is the list of columns this dataset has:

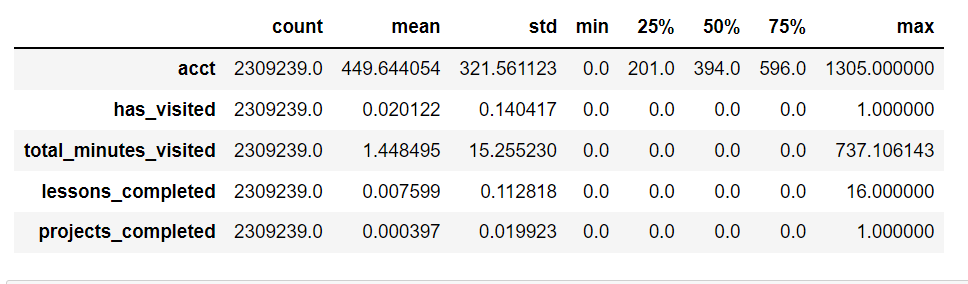


**Table 2 - Columns in Dataset**

1. **EDA & Business Implication**

EDA stands for exploratory data analysis where we explore our data and grab insights from it. EDA helps us in getting knowledge in form of various plots and diagrams where we can easily understand the data and its features.

**Analysis Of the Data**



**Table 3- Checking of the Skewness in Data**

**Observation:**

**acct:** account key for each user value is in 0 - 1305 range. As Mean is greater than Median data is rightly skewed.

**has\_visited:** the number of visits is in 0 – 1 range. As Mean ~ Median it’s

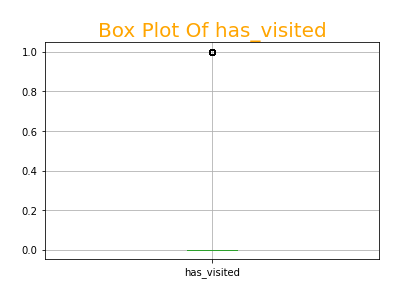
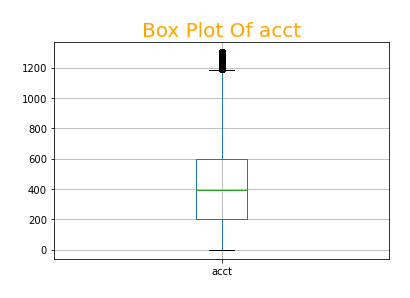
almost Normal Distributed.

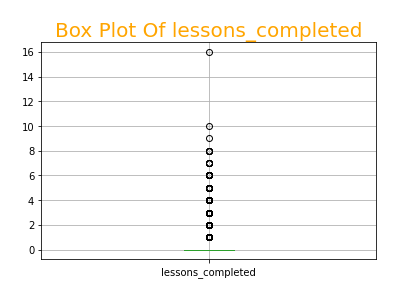
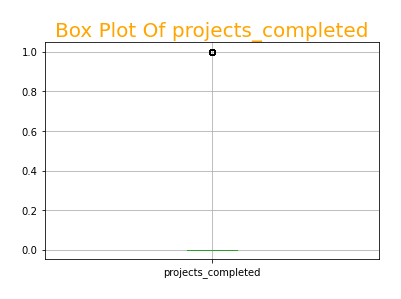
**total\_minutes\_visited:** the number of visits is in 0 – 737 range. As Mean is greater than Median data is rightly skewed.

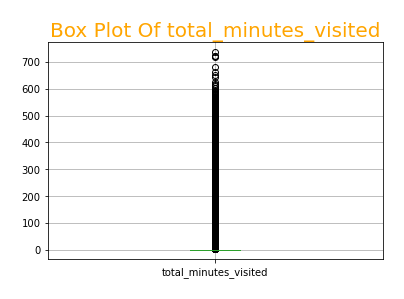
**lessons\_completed:** the number of competed lessons is in 0 – 16 range. As Mean ~ Median it’s almost Normal Distributed.

**projects\_completed:** the number of competed projects is in 0 – 1 range. As Mean ~ Median it’s almost Normal Distributed.

**Uni-variate Analysis - Box Plot**







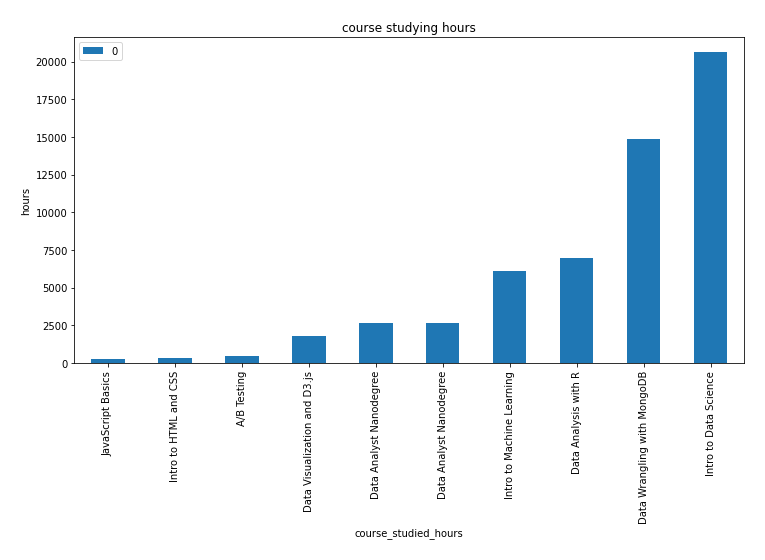
**Plot 1- Boxplot**

**Observation:**

Most of features are having outliers**.**

**Uni-Variate Analysis of Important relations – Bar Plot**

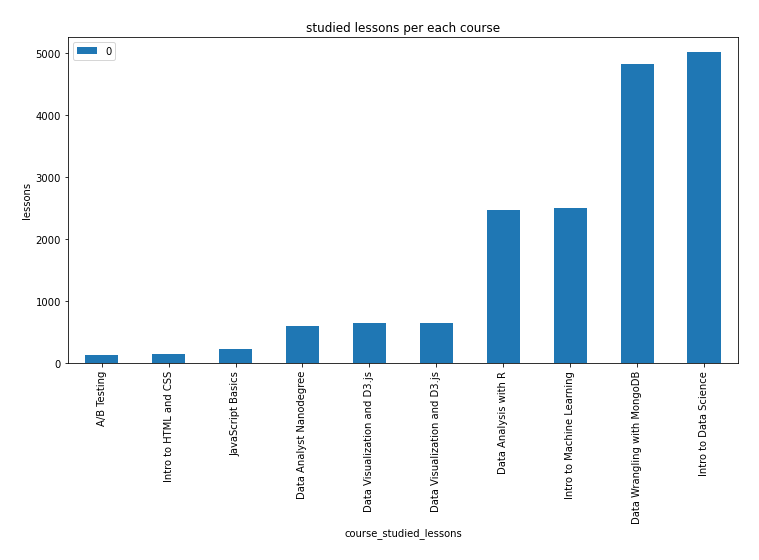
* Analysis of time spent for studying every course:



**Studying time per course**

According to studying hours, “Intro to Data Science” is the most studied course.

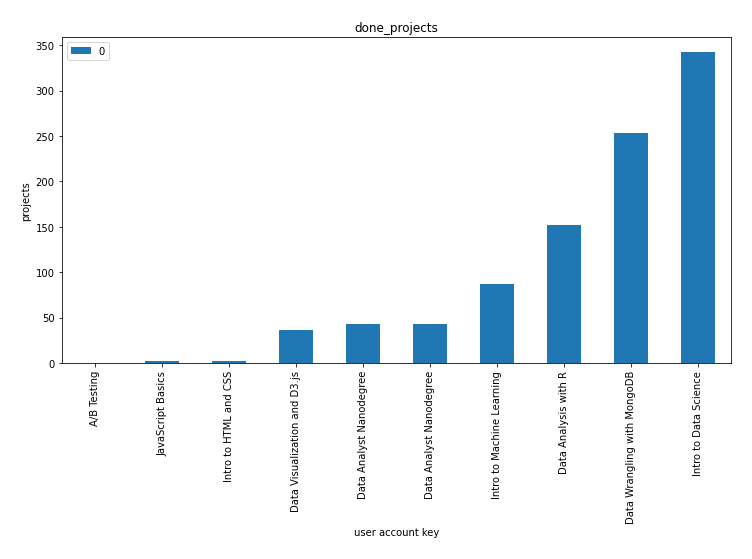
* Analysis of lessons completed for every course:



**completed lessons per course**

According to lessons done, “Intro to Data Science” is the most studied course.

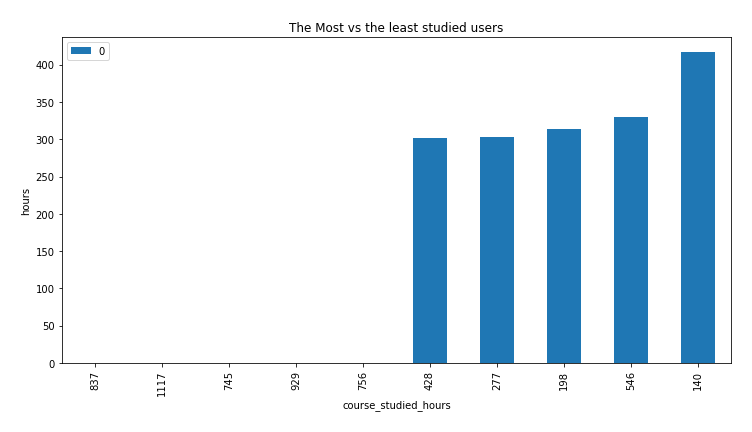
* Analysis of projects completed for every course:



**Completed projects per course**

According to projects done, “Intro to Data Science” is the most studied course.

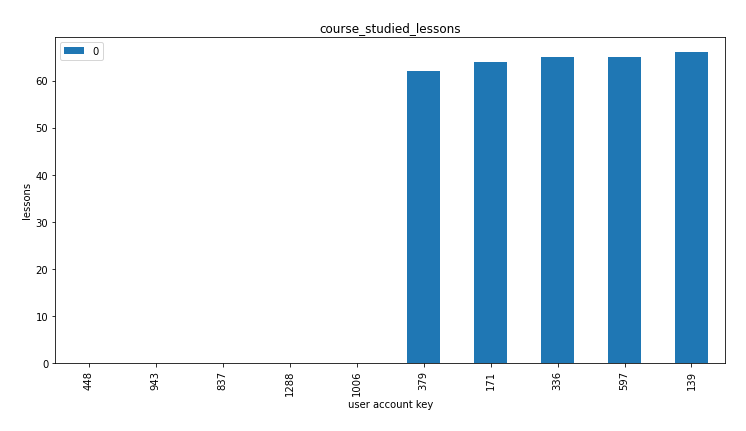
* “Intro to Data Science” course is the most popular course for all users.
* Analysis of time spent for studying per user:



**Studying time per user**

According to studying hours, user whose account key equals to 140 is the most user studied.

* Analysis of lessons completed for every course:

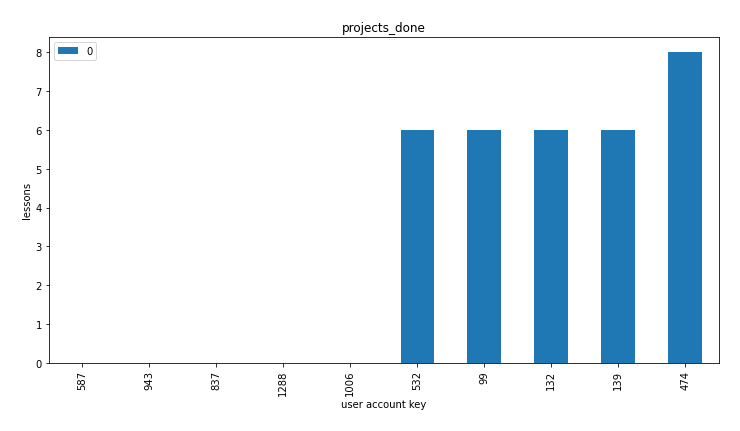


**Lessons completed per user**

**Plot 2- Barplot**

According to lessons done, user whose account key equals to 139 is the most user done lessons.

* Analysis of projects completed for every course:



**Project done per user**

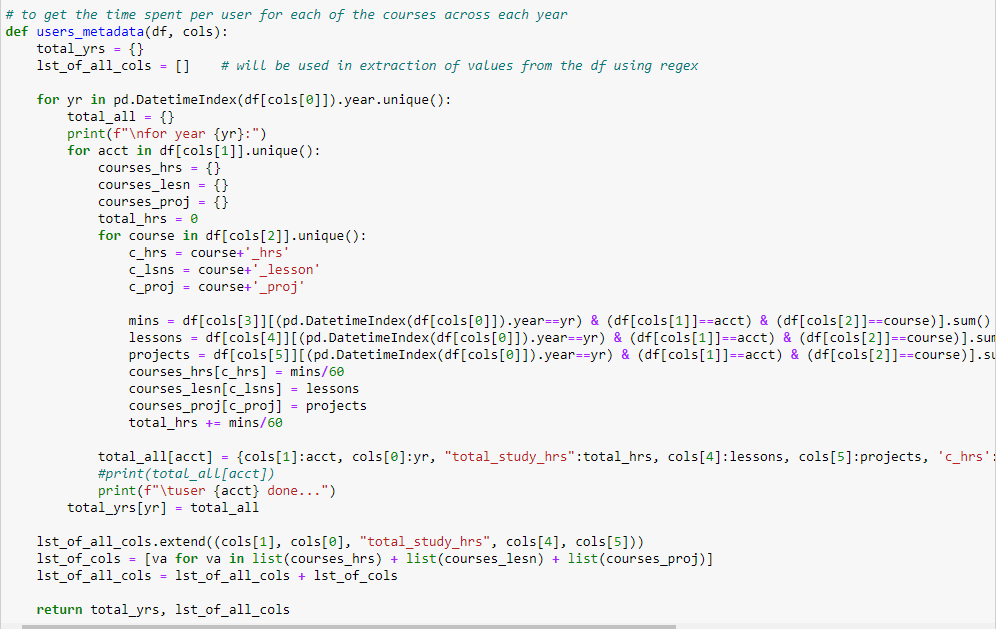
According to projects done, user whose account key equals to 474 is the most user applied projects.

* Users whose account key equals 140, 139 and 474 are special users for Udacity platform, which allows them to take more discounts on courses they studying.

1. **User progress per each year**

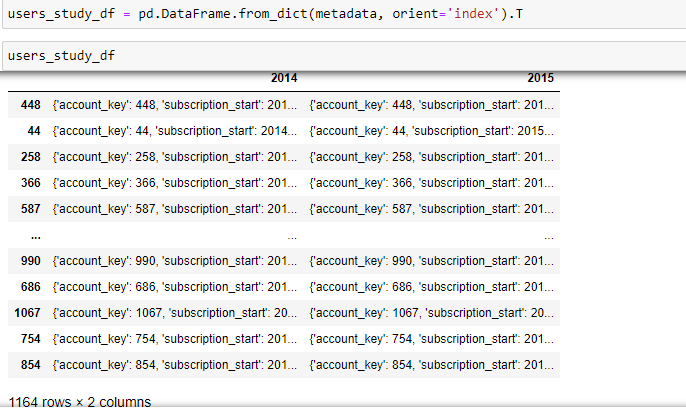
To get the details of each student in order to meet their needs it’s good to calculate the studied hours , lessons and projects per each course of the courses we have in the “course title”.

The code below collects each student data into a dictionary and creates a nested dictionary contains each year as a key and the corresponding data in detailed about every user.



**Metadata**

Thenwe convert the output from a dictionary to a dataframe:

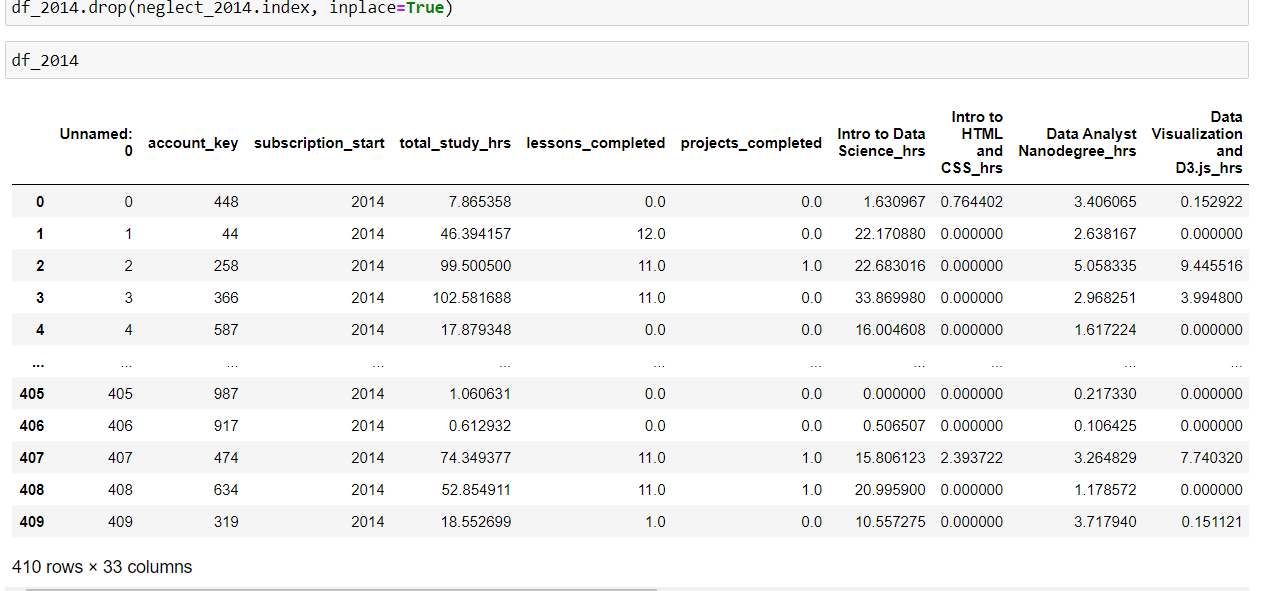


Then we pass the created dataframe “users\_study\_df” to the following function to clean this messy form of data and returns the important user information for us

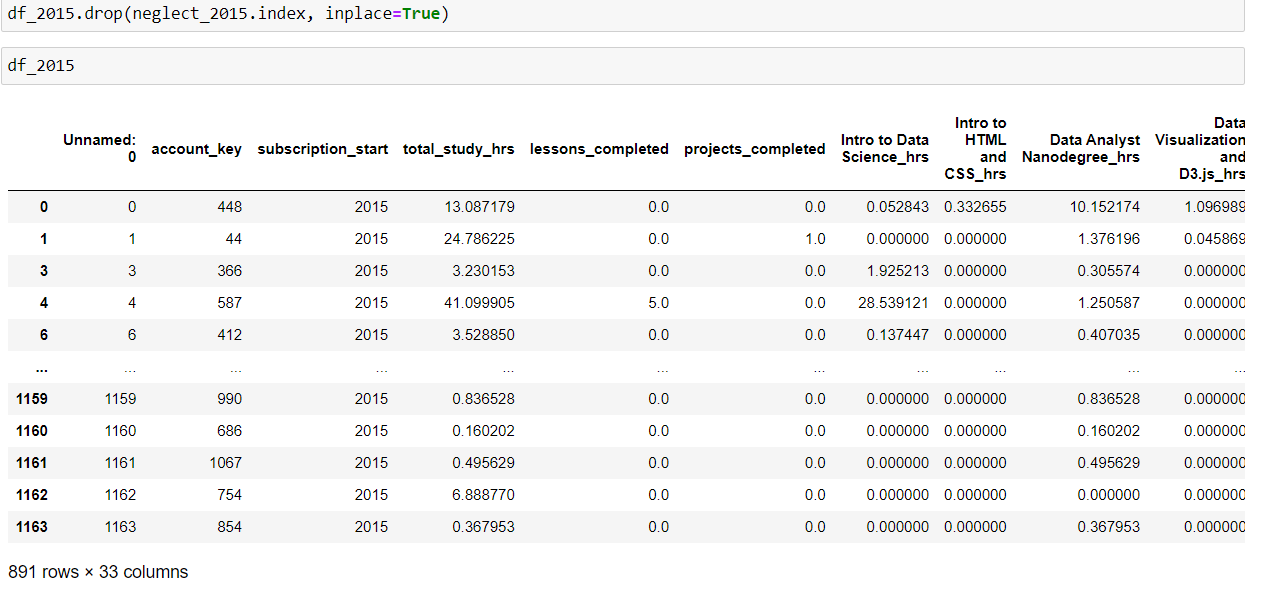
**Regex for cleaning data**

We now saved our cleaned data for each year which describes all important information about each user.

*Actually there's a little problem, which is the csv files that are created for students contains all of them for both years, and that's not true, it's not a real problem as the data filled for those students that didn't achieve any progress during the year as zeros for all columns, but if we remove them we'll have a more cleaned data, so it's optional, and I'm going to remove it and save my data once again without any overwriting*

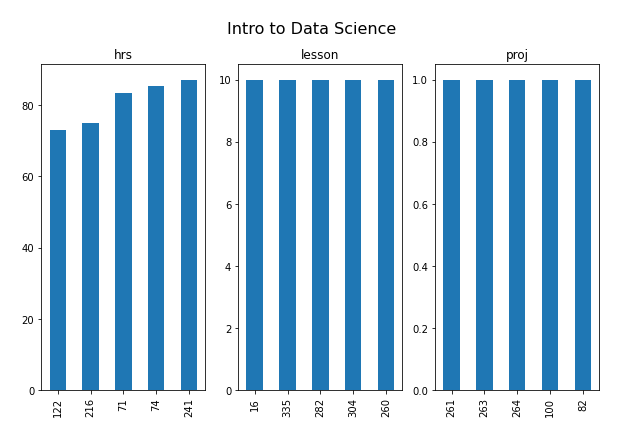
Data after cleaning: 

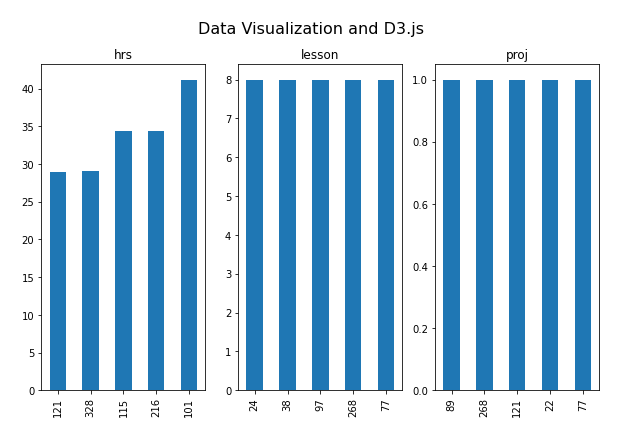
**Data for 2014**

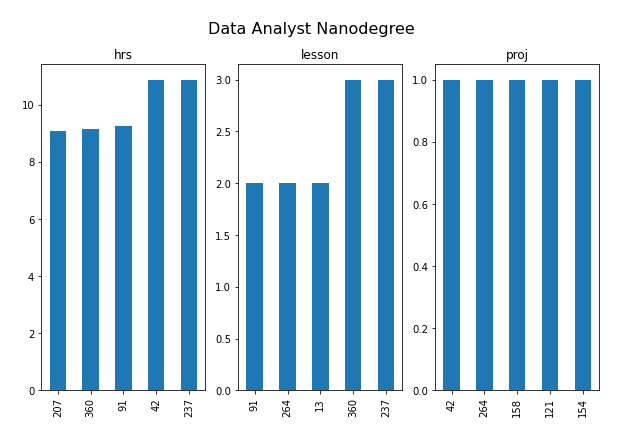
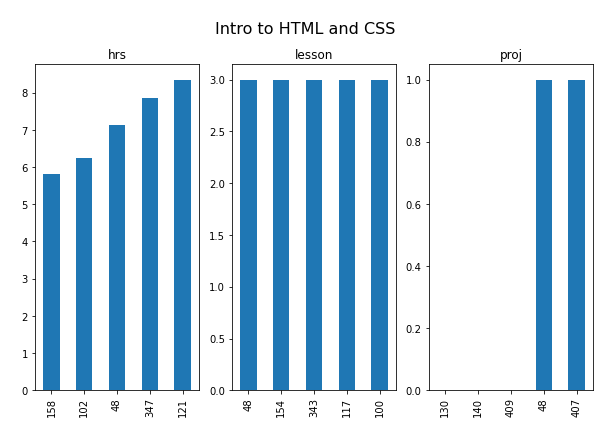
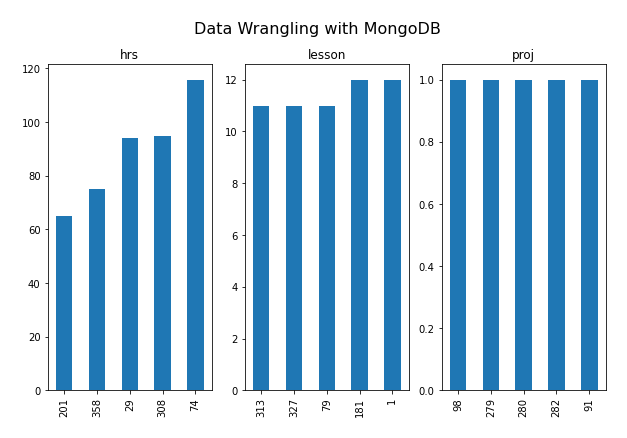
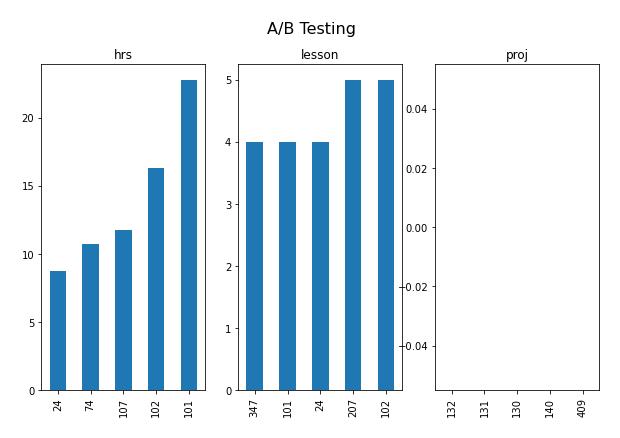
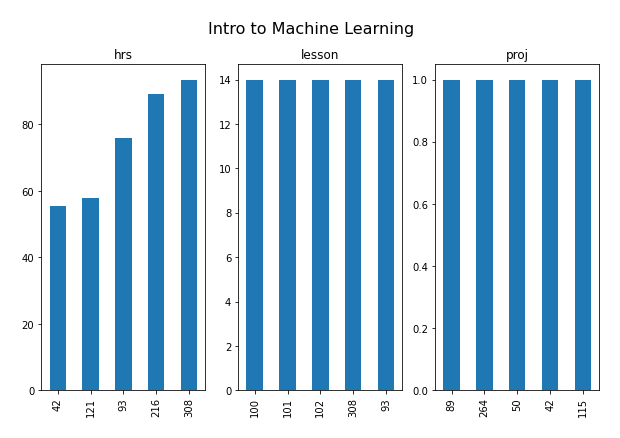
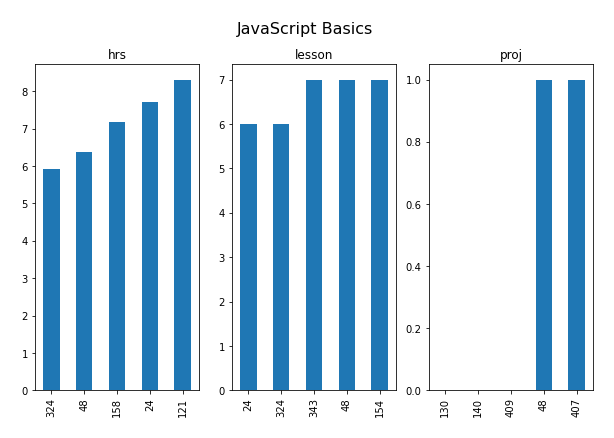
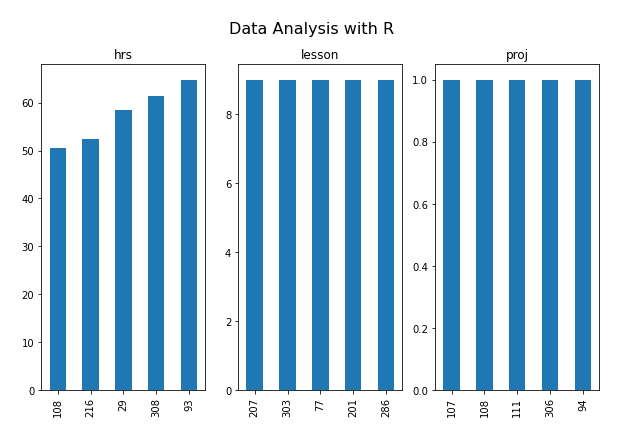


**Data for 2015**

**Uni-Variate Analysis of Important relations for year 2014 – Bar Plot**

****





**Final Words:**

We now contains a cleaned dataset for the user progress per year which can help us for clustering our users into categories if we applied a ML model for that.