DOUBT RESOLUTION SYSTEM ANALYSIS

ANJALI CHOPRA

AGENDA

- Introduction
- Descriptive Analysis
- Areas of Improvement
- Executive Summary
- Annexure

INTRODUCTION:

Performing Data Analysis on Coding Ninja's Doubt Resolution System in order to gain meaningful insights from past records to improve the overall performance of all the courses offered by this Ed-Tech Platform.

DESCRIPTIVE ANALYSIS

- Descriptive analysis is the process of using current and historical data to identify trends and relationships.
- Doubt Resolution data for the year 2020 was taken to perform an in-depth Analysis.

TASK 1: Overall analysis around the resolution time

- Merged two excel-worksheets based on a shared column, i.e., "id" in the doubts sheet & "doubt_id" in the activities sheet using VLOOKUP Formula.

Formula : = VLOOKUP(A15,'https://d.docs.live.net/4dd91e3962adde37/Desktop/Coding ninjas data analyst/data/[updated activities.xlsx]activities'!$C:$E,{2,3},0)

- Analyzed overall resolution time using the two columns "Doubt_Table_time" & "Activity_Table_Time".

- Data was present in YYYY-DD-MM HH:MM:SS format in the above-mentioned columns.

- Segregated the time format from the UTC date time format for further Analysis using the delimiter functionality of Excel.

- As a part of feature engineering, two new columns are created:
  1). Doubt_Creation_Time
  2). Doubt_Resolved_Time

Short key: Alt + D + E + D, then delimited the data separated by space and extracted the time from the given Date-Time format.

- Doubt Resolution Time: It denotes the difference in time at which a particular doubt is raised and resolved by a TA.

- It is calculated by subtracting the doubt creation time from the time of 'resolve' activity on the doubt.

Short key: Alt + D + F + F, to filter the results based on the 'Resolve' activity type.

Formula: = ABS(Doubt_Resolved_Time) - Doubt_Creation_Time)

- ABS: This functionality gives the Absolute Difference between the two-time stamps by ignoring the sign of the value. It will prevent the AM/PM difference errors.

- A Histogram is used to depict the Relationship Between the Average Doubt Count V/S Doubt Resolution Time.

Task 2: Analysis of the effect of Resolution Time on the User Rating

- Extracted two continuous columns i.e., “Doubt_Resolution_Time” & “user_rating” from the Dataset to Perform Bivariate Analysis on the given feature set.

- Handled a single missing value in the user_rating column & imputed the blank cell with the mean of the mentioned column.

- Computed Average of the given range of user_rating column and Rounded OFF up to 0 decimal place in order to get an exact integer value that comes out to be **4**.

Formula : = ROUND(AVERAGE(B6:B236),0)

- Analyzed the impact of Doubt Resolution Time on User Rating using the Bivariate Analysis Methodology for continuous variables i.e., **“Correlation Coefficient”.**

- It is calculated by using the “correl” function applied to the range of 2 columns.

Formula : = CORREL(A6:A236,B6:B236)

- Correlation coefficient is used to measure the strength and direction of a linear relationship between two variables. Its value of correlation coefficient ranges from -1 to +1.

OBSERVATION

- Closer to +1: A coefficient value of 1 represents a **Perfect Positive Correlation**. The closer the coefficient is to +1, the higher the correlation. As our Independent variable increases, the Dependent variable increases as well.

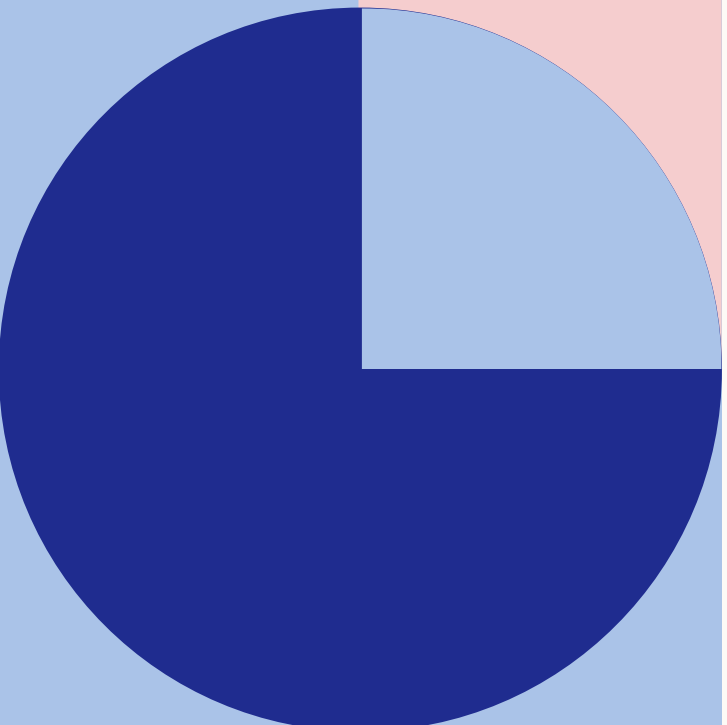
- Closer to -1: A coefficient of the value of -1 represents a **Perfect Negative Correlation**. The closer the coefficient is to -1, the lower the correlation. As our independent variable increases, the Dependent variable decreases.

- Value equal to 0: A correlation coefficient near 0 indicates no correlation, which means there is **No Particular Relationship** between the variables.

Applied Correlation Function on the two continuous columns i.e. Doubt_resolution_Time & user_rating and the **Correlation Coefficient** comes out to be **-0.08** which clearly states that both columns are **Negatively Correlated**. So It is observed that as the Doubt Resolution Time **Increases**, the user Rating **Decreases**.

TASK-3: HOW MANY DOUBTS ARE RESOLVED BY EACH TA?

A Pivot Table is a powerful tool to calculate, summarize, and analyze data that lets you see comparisons, patterns, and trends in your data.



- **In this Task, A pivot table is used to summarize the Count of Resolved Doubts by each Teaching Assistant. User_id is added to the 'rows' pane and count functionality is applied over the user_id & added to the "values" pane in the pivot table and the data is sorted in descending order.**
- **Doubt Activity: Every time either the user, TA or the system does an activity we create an entry in the Doubt Activity table.**
- **User_id: It denotes the unique Identification of either a TA, student or the system. User_id will be the id of the TA in case of the following activity: activate, reject, pending_action_required, pending_information_required, review_resolution, assign, accept, reactivate, and escalate.**
- **We already filtered out the 'resolve' activity state in the merged worksheet in order to get exclusively the user_id of TAs.**
- **It is observed that the maximum doubts are taken by user_id "398570" and the doubt count is "42".**

TASK 4: Overall Performance of Doubt Resolution in Different Courses

- An Interactive dashboard is created via Pivot Table and Slicers in Excel to showcase the entire Doubt Resolution Performance of different courses.
- Different parameters are added to multiple fields in the pivot table to analyze the result.
- The **"Activity Type"** is added to the **Filters pane** to filter out the result based on various stages of the doubt-resolution process, it is sub-categorized as **'activate', 'resolve' & 'assign'.**
- All the courses identified by their **"course_ids"** are added to the **Rows pane** in order to visualize the performance of different courses with other parameters.
- The **COUNT** operation is applied over the **User_rating** to get the count of the overall rating received by each course.
- The most significant parameter i.e., **"User Rating"** is Added to the **Columns pane** in order to visualize the rating corresponding to each course.
- A ***SLICER*** of **"Content-Type"** is added to analyze the user_rating received for different courses based on the content of the doubt raised by a student.

TASK 5: AREAS OF IMPROVEMENT

DOUBT RESOLUTION TIME

In order to improve the overall performance of a course, every TA should accept the doubts earlier in order to get rid of fewer ratings.

ASSIGN TO ACCEPT RATIO

TAs should accept more doubts on a daily basis in order to improve the performance of the course. The assign-to-accept ratio should be on the lower level.

FOCUS ON CONTENT TYPE

Mostly Students asked Coding Related problems so TAs should Take maximum Coding Problems and Provide Proper Resolution to the Students.

EXECUTIVE SUMMARY

- Based on the Descriptive Analysis of the year 2020 Dataset, it is analyzed that **course_id 1,** is the top performer among the other courses as it holds **1313** a 5-star rating out of **3761.** & its Grand Total rating is also the highest among the other course_ids i.e, 1725.

- **Course_id 21**, needs maximum improvement in terms of overall rating as it received the least rating count among the other courses.

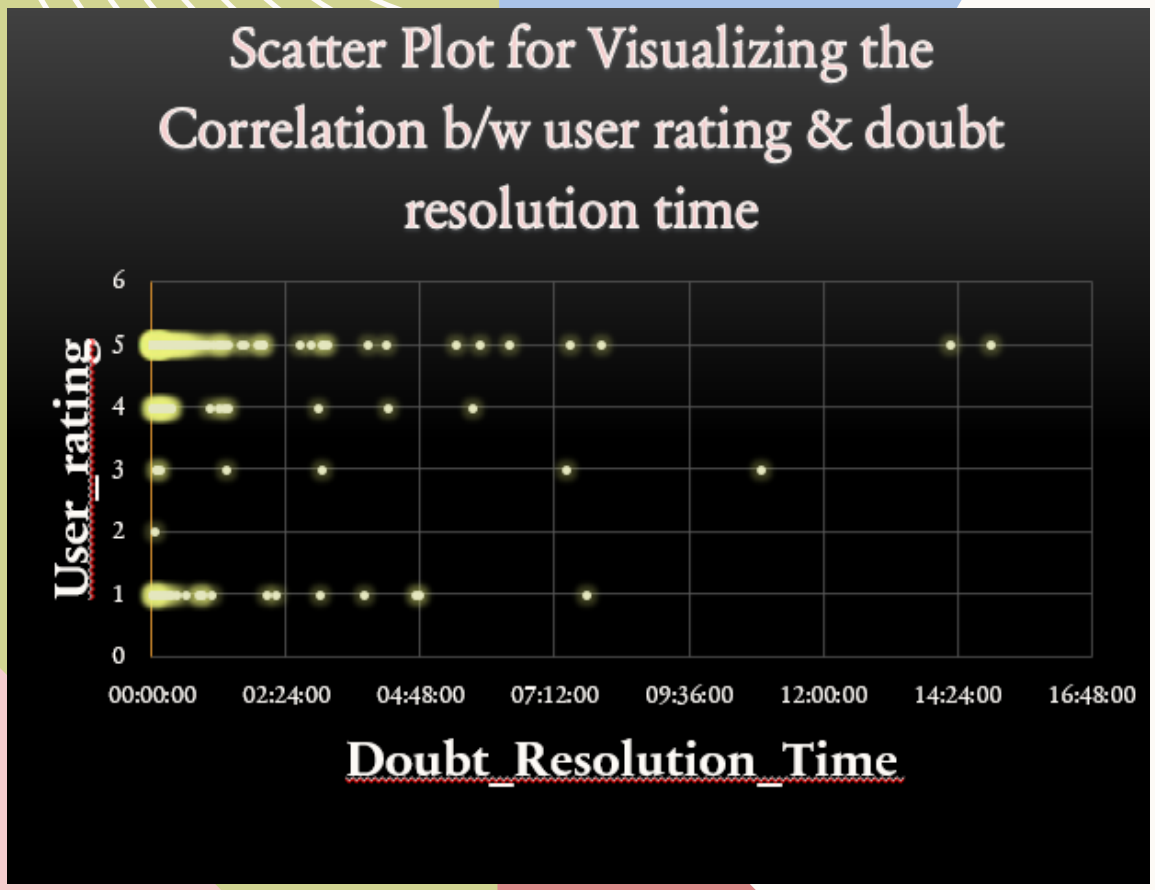
- It is Observed that **the Doubt Resolution Time** affects the **user rating** as both the continuous variables are **negatively correlated** so an increment in the resolution time will degrade the rating.

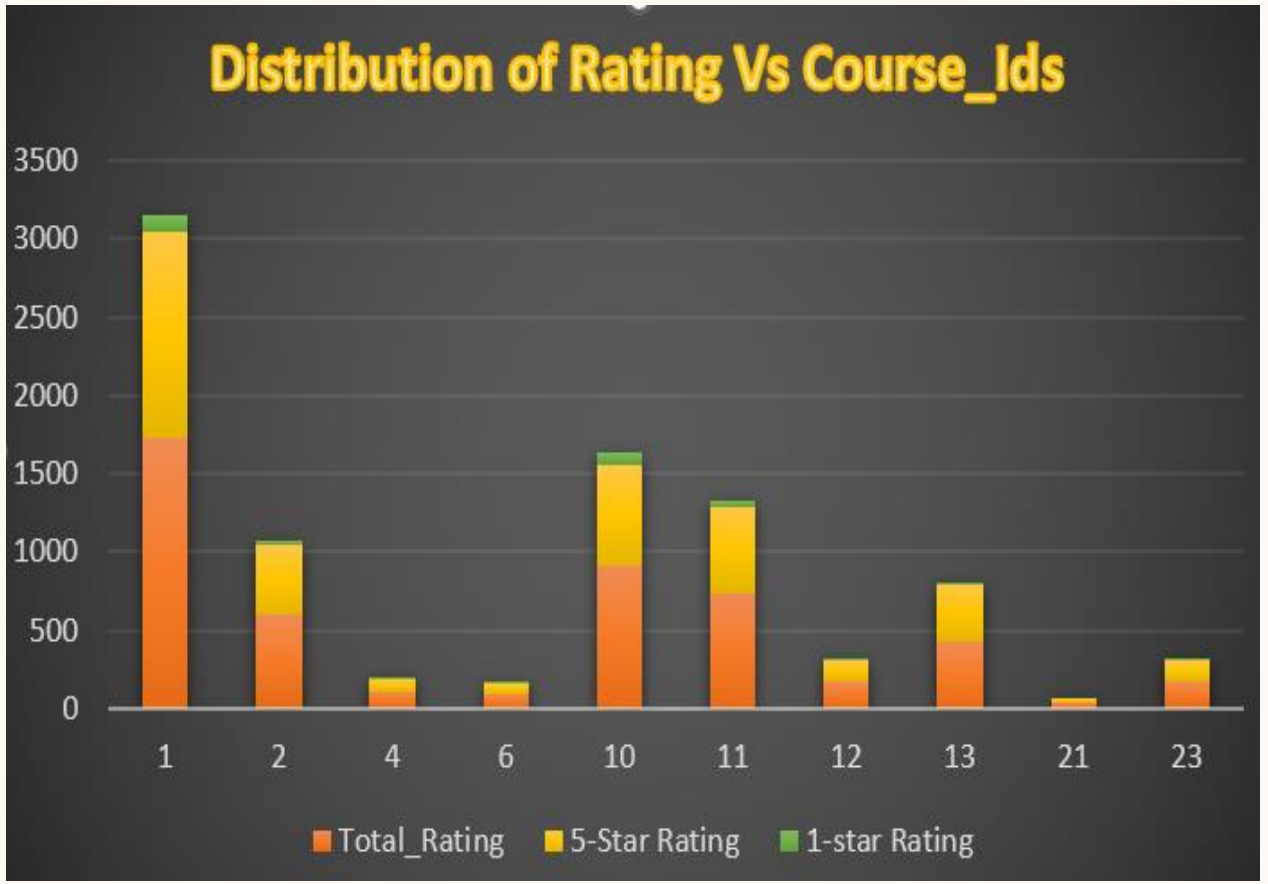
- **"Coding problems"** is the highest asked **Content Type** by students followed by **"Single Choice Problem"** as per observation.

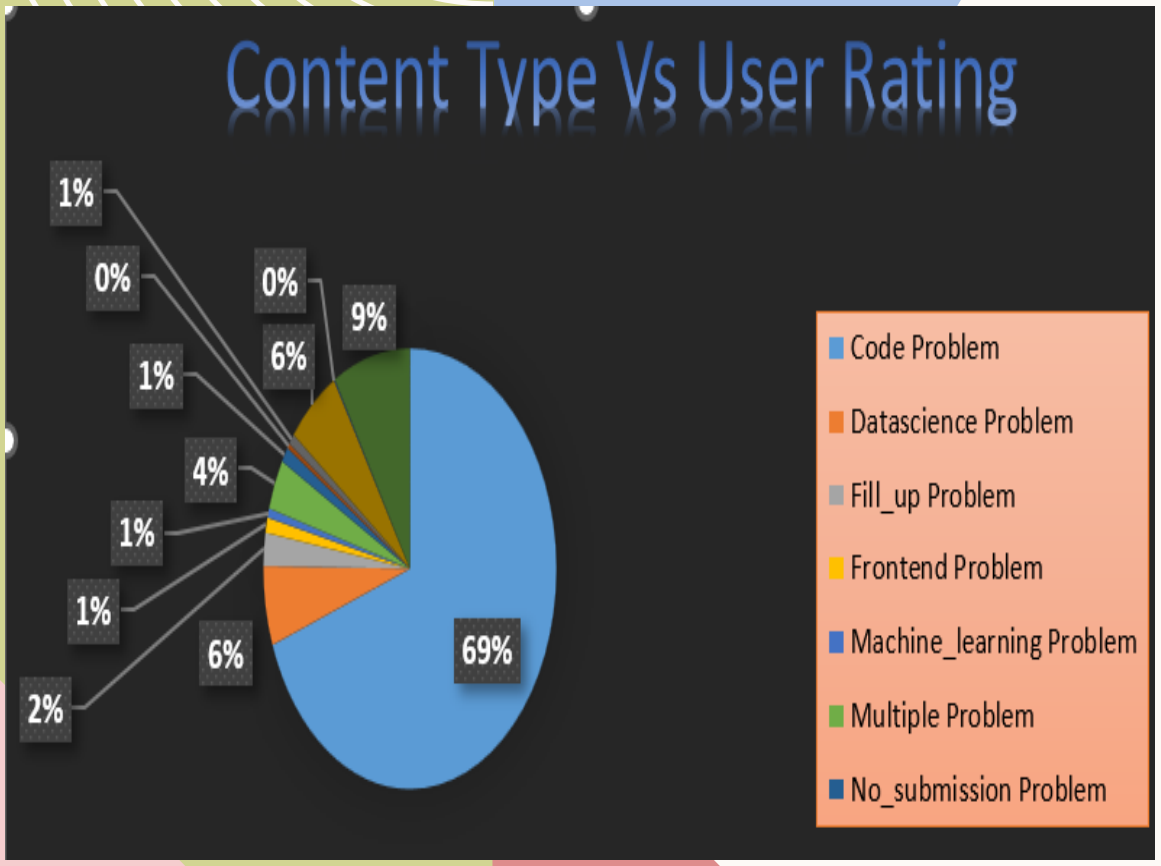
- It is explored that the maximum doubts are taken **by user_id "398570"** and the **doubt count** is **"42"** and might get the best compensation out of other Teaching Assistants**.**

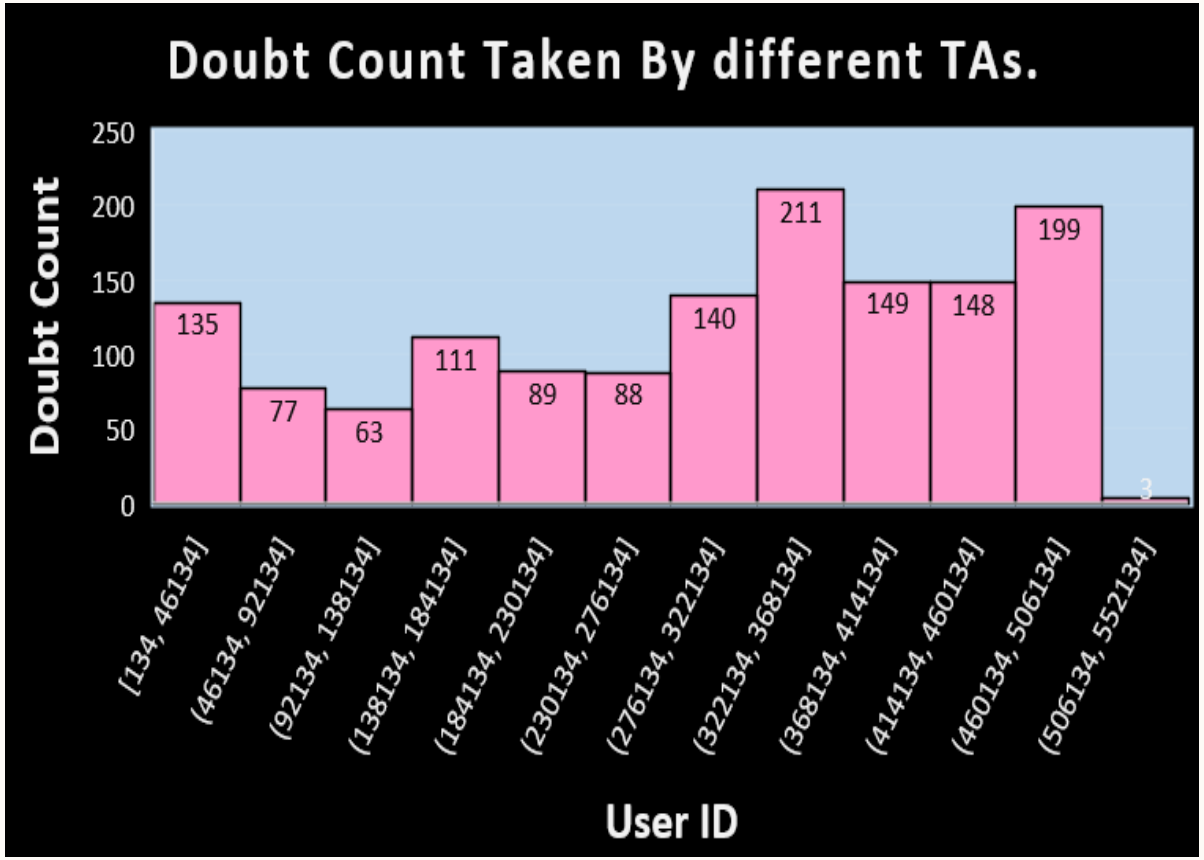
- Assign to accept ratio on average is **very high** as lesser doubts are resolved in comparison to the assigned ratio.

ANNEXURE


Scatter Plot for Visualizing the Correlation b/w user rating & doubt resolution time
User_rating
6
5
4
3
2
1
0
00:00:00
02:24:00
04:48:00
07:12:00
09:36:00
12:00:00
14:24:00
16:48:00
Doubt_Resolution_Time


Distribution of Rating Vs Course_Ids
3500
3000
2500
2000
1500
1000
500
0
1
2
4
6
10
11
12
13
21
23
Total_Rating
5-Star Rating
1-star Rating


Content Type Vs User Rating
1%
0%
0%
9%
6%
1%
4%
1%
1%
6%
2%
69%
Code Problem
Datascience Problem
Fill_up Problem
Frontend Problem
Machine_learning Problem
Multiple Problem
No_submission Problem


Doubt Count Taken By different TAs.
Doubt Count
250
200
150
100
50
0
135
77
63
111
89
88
140
211
149
148
199
3
[134, 46134]
(46134, 92134]
(92134, 138134]
(138134, 184134]
(184134, 230134]
(230134, 276134]
(276134, 322134]
(322134, 368134]
(368134, 414134]
(414134, 460134]
(460134, 506134]
(506134, 552134]
User ID

THANK YOU