Student Performance Data Set

Source:
https://www.kaggle.com/uciml/student-alcohol-consumption
https://archive.ics.uci.edu/ml/datasets/student+performance

Dataset descriptions:

- Attributes = 28 (number 1 - 28)
- Measure = 5 (number 29 - 33)
- Aggregate Functions = Max, Sum, Avg, Std

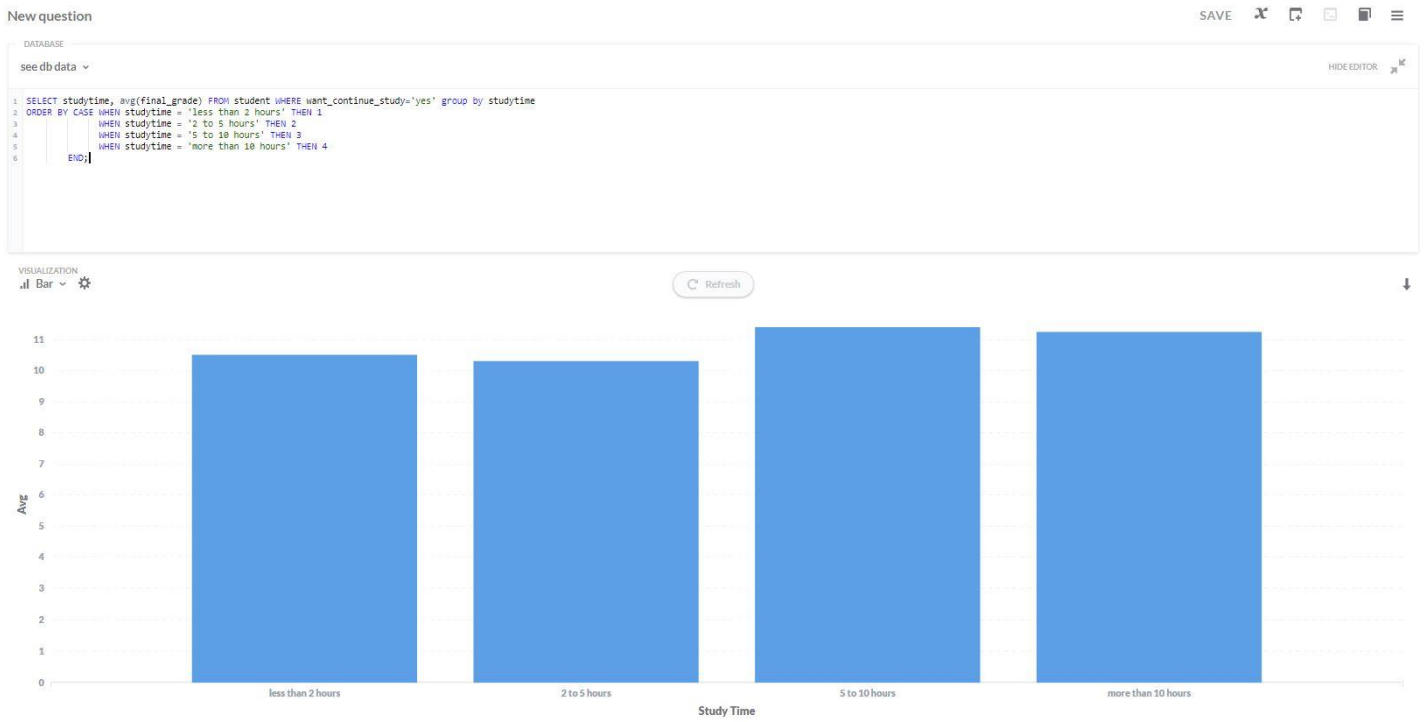
The data were obtained in a survey of students math and Portuguese language courses in secondary school. It contains a lot of interesting social, gender and study information about students. In this experiment, only math dataset is used. Detail of the attributes and measure as follows:

1. school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2. sex - student's sex (binary: 'F' - female or 'M' - male)
3. address - student's home address type (binary: 'U' - urban or 'R' - rural)
4. famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
5. Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
6. Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 -“ 5th to 9th grade, 3 -“ secondary education or 4 -“ higher education)
7. Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 -“ 5th to 9th grade, 3 -“ secondary education or 4 -“ higher education)
8. Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
9. Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10. reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
11. guardian - student's guardian (nominal: 'mother', 'father' or 'other')
12. traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
13. studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
14. failures - number of past class failures (numeric: n if 1<=n<3, else 4)
15. schoolsup - extra educational support (binary: yes or no)
16. famsup - family educational support (binary: yes or no)
17. paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
18. activities - extra-curricular activities (binary: yes or no)
19. nursery - attended nursery school (binary: yes or no)
20. higher - wants to take higher education (binary: yes or no)
21. internet - Internet access at home (binary: yes or no)
22. romantic - with a romantic relationship (binary: yes or no)
23. famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
24. freetime - free time after school (numeric: from 1 - very low to 5 - very high)
25. goout - going out with friends (numeric: from 1 - very low to 5 - very high)
26. Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
27. Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
28. health - current health status (numeric: from 1 - very bad to 5 - very good)

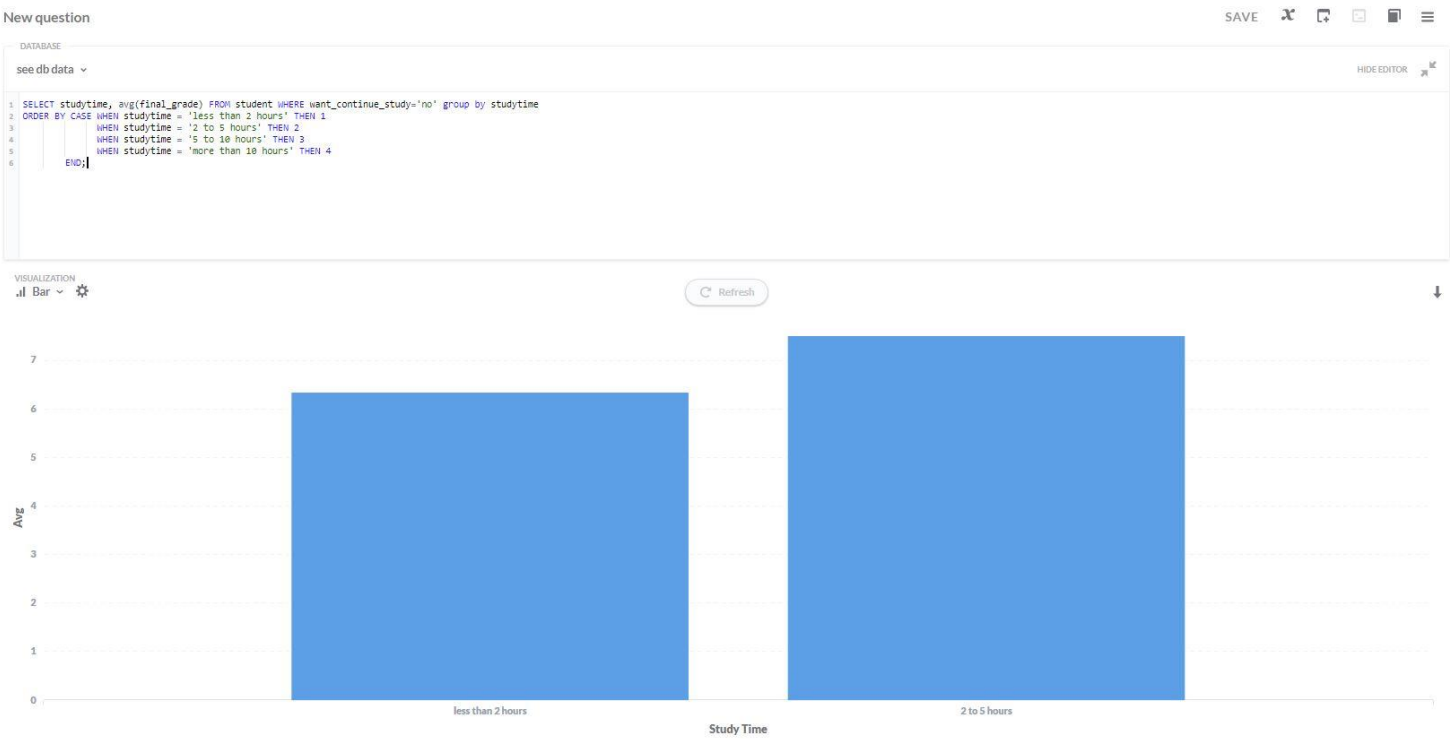
29. absences - number of school absences (numeric: from 0 to 93)
30. age - student's age (numeric: from 15 to 22)
31. G1 - first period grade (numeric: from 0 to 20)
32. G2 - second period grade (numeric: from 0 to 20)
33. G3 - final grade (numeric: from 0 to 20, output target)

For instance, the analyst want to compare between students who want to continue their study to the higher education and students who do not want to continue their study to higher education.

Importance score: *0.812831582477977*

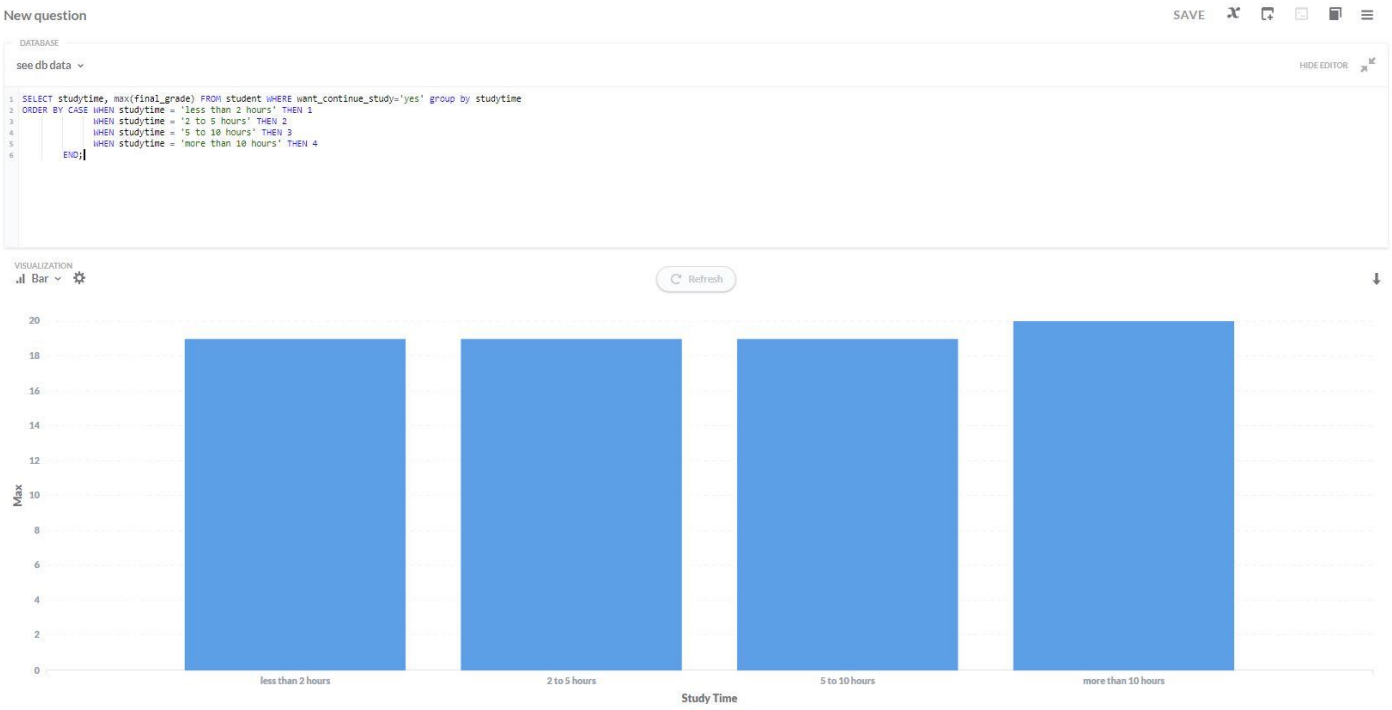


Target query: *SELECT studytime,* ***AVG****(final_grade) FROM student WHERE want_continue_study='****yes****' group by studytime*

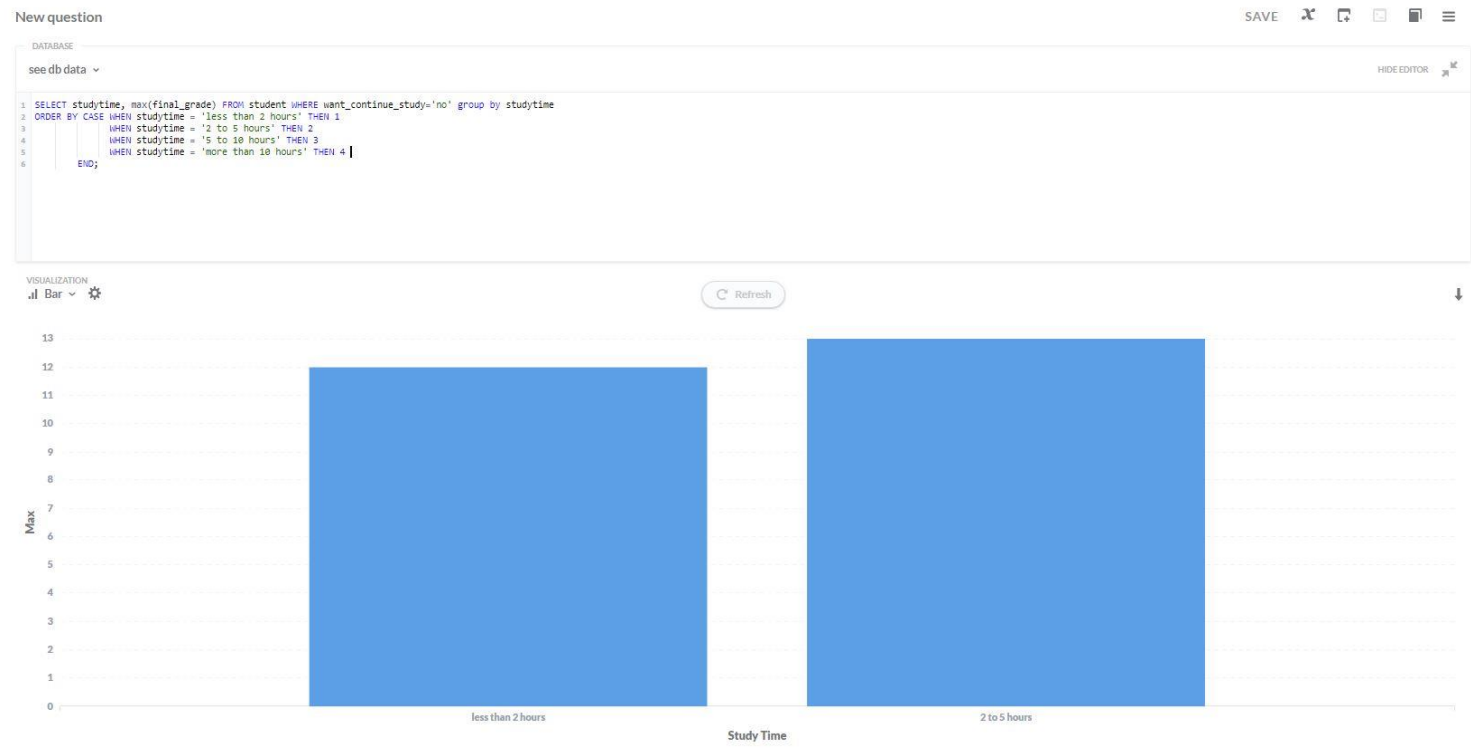


Reference query: *SELECT studytime,* ***AVG****(final_grade) FROM student WHERE want_continue_study='****no****' group by studytime*

***Importance score**: 0.797221012363291*



Target query: *SELECT studytime, **MAX**(final_grade) FROM student WHERE want_continue_study='**yes**' group by studytime*



Reference query: *SELECT studytime, **MAX**(final_grade) FROM student WHERE want_continue_study='**no**' group by studytime*

As shown in the Figure above, students who want to continue to higher education relatively have better final grade compared to students who do not want to continue to higher education. Interestingly, some students who want to continue their education, they spend more time to study (5 to 10 hours, even more than 10 hours) per week. To contrary, students who do not want to continue their study, all of them only spend less than 5 hours a week for study.

Kickstarter Dataset

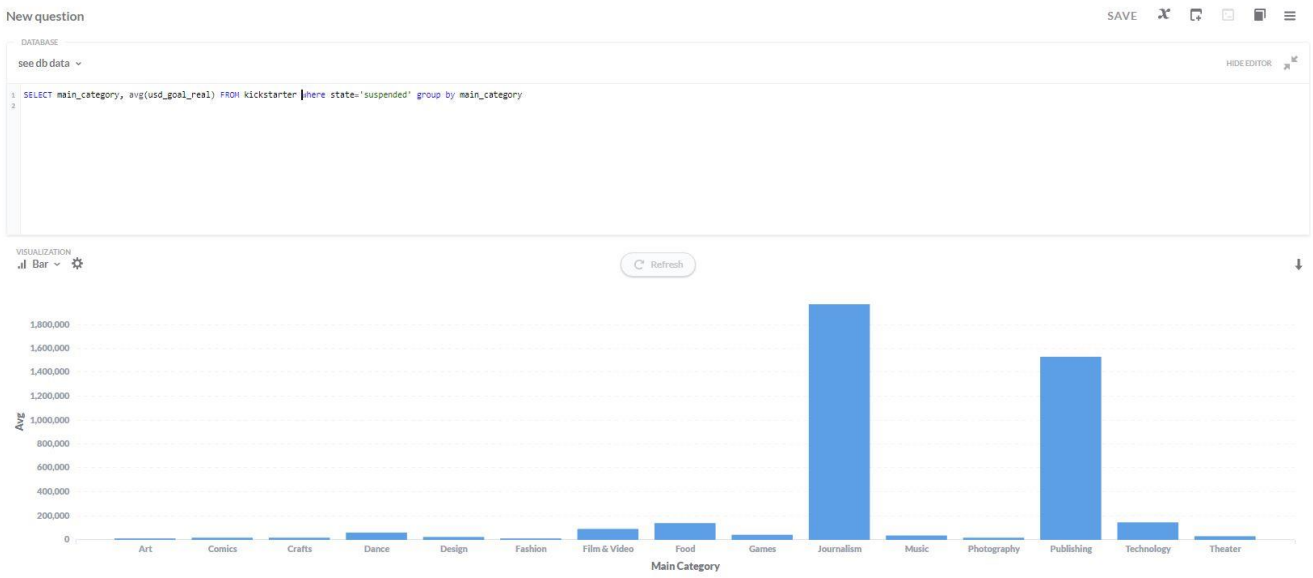
Source: https://www.kaggle.com/kemical/kickstarter-projects
Dataset descriptions:

- Attributes = main_category (e.g., Art, Technology), state (e.g., successful, failed, suspended), country (e.g., USA, Singapore, UK)
- Measure = backers (i.e., project supporters), used_pledged_real (pledged amount of money from backers), used_goal_real (amount of money for the project in total)
- Aggregate Functions = Max, Sum, Avg, Std

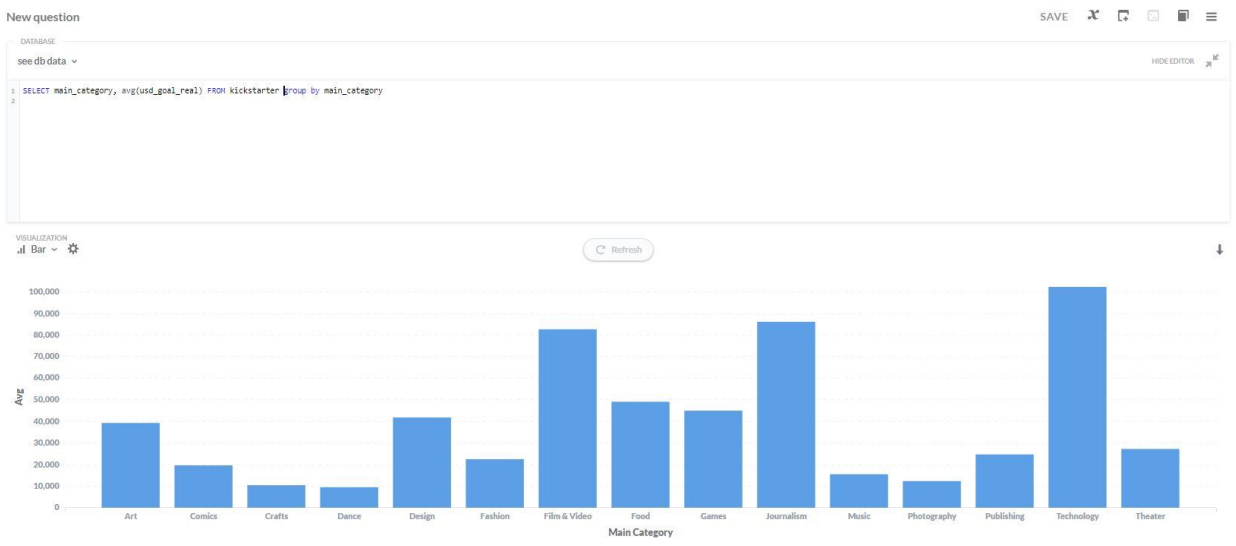
While I include attribute 'country' as I expected that the top-k views are dominated by this attribute. It's because if 'country' attribute is used as *X* axis then its view may has high deviation compared to the reference subset due to a lot of bars are generated and some of them may missing on the reference subset. To overcome this issue, 'country' attribute is not included.

In this experiment, I compared all subsets from this dataset to whole dataset and find the top-k views which has highest importance score and here the result:

Importance score: 0.519404384759085



Target query: *SELECT main_category, **AVG**(usd_goal_real) FROM kickstarter WHERE state='suspended' GROUP BY main_category*

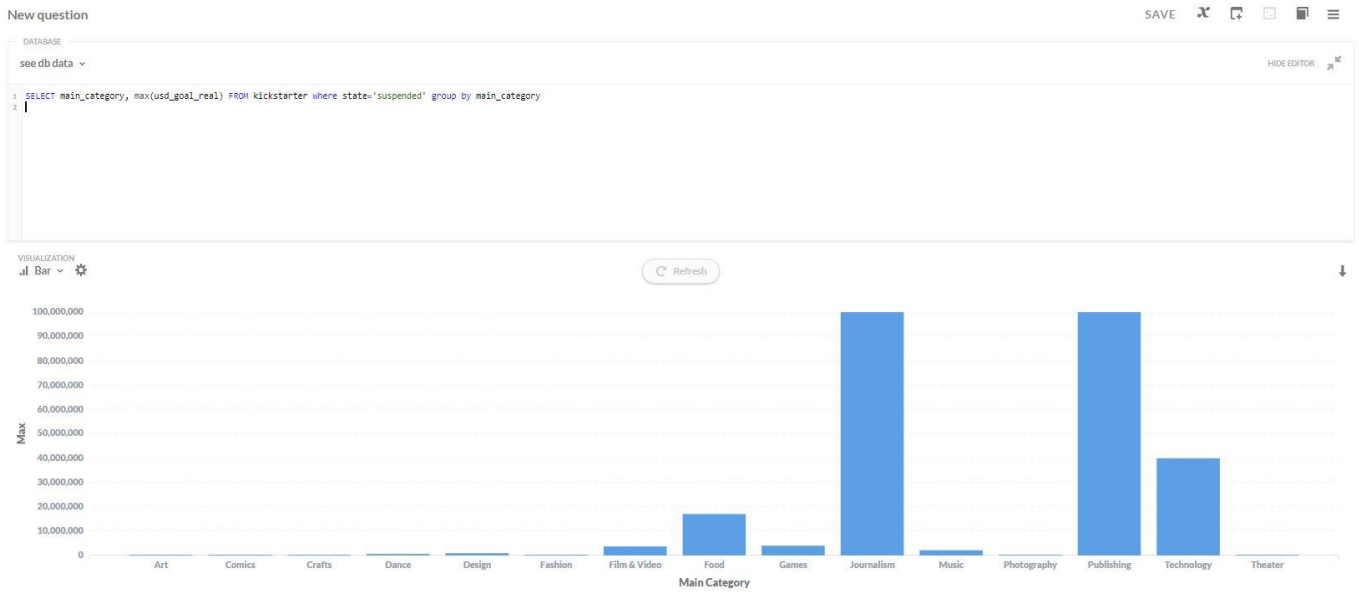


Reference query: *SELECT main_category, **AVG**(usd_goal_real) FROM kickstarter GROUP BY main_category*

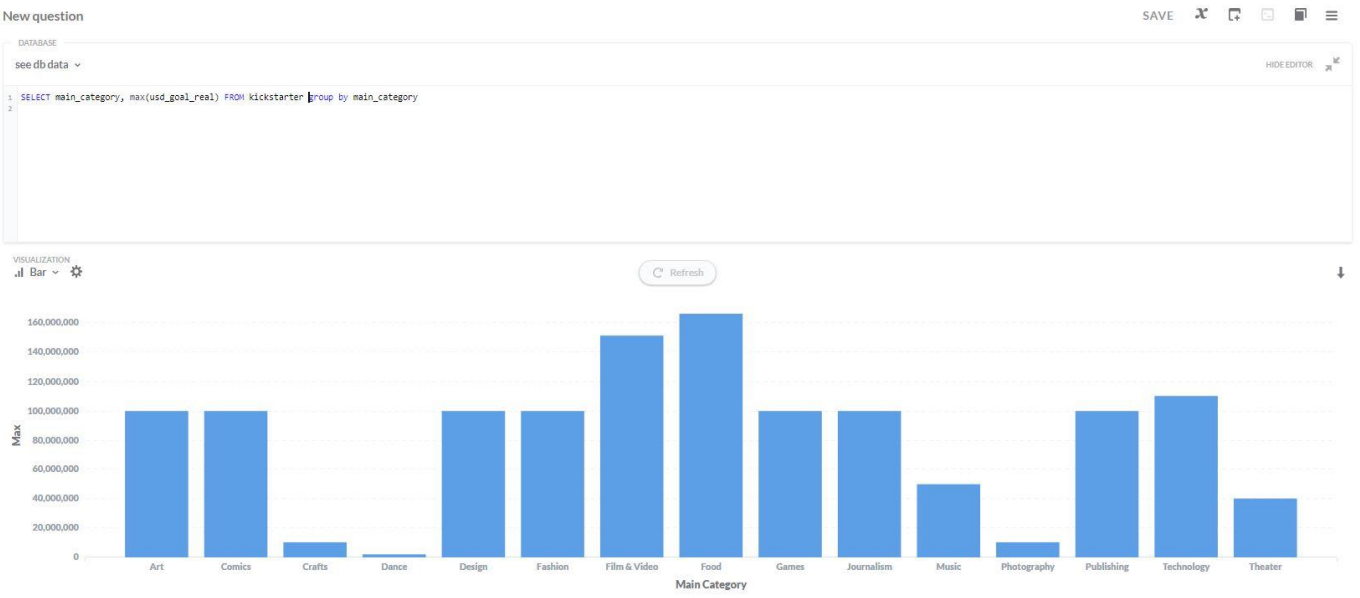
That view shows the amount of money needed to do the project as Y axis and the project categories in the Kickstarter's website as X axis. As shown in that Figure, if we look at the reference view, the highest average money needed for the project is owned by Technology category (i.e., around 100,000), Journalism, Film & Video category (i.e., under 90,000) and other categories are below of them.

However, if we see the 'suspended' subset (i.e., project that has suspended by Kickstarter), the highest average money needed for the project is owned by Journalism category (i.e., above 1,800,000) and Publishing category (i.e., around 1,500,000).

Importance score : 0.470106628498069



Target query: *SELECT main_category, **MAX**(usd_goal_real) FROM kickstarter WHERE state='suspended' GROUP BY main_category*



Reference query: *SELECT main_category, **MAX**(usd_goal_real) FROM kickstarter GROUP BY main_category*