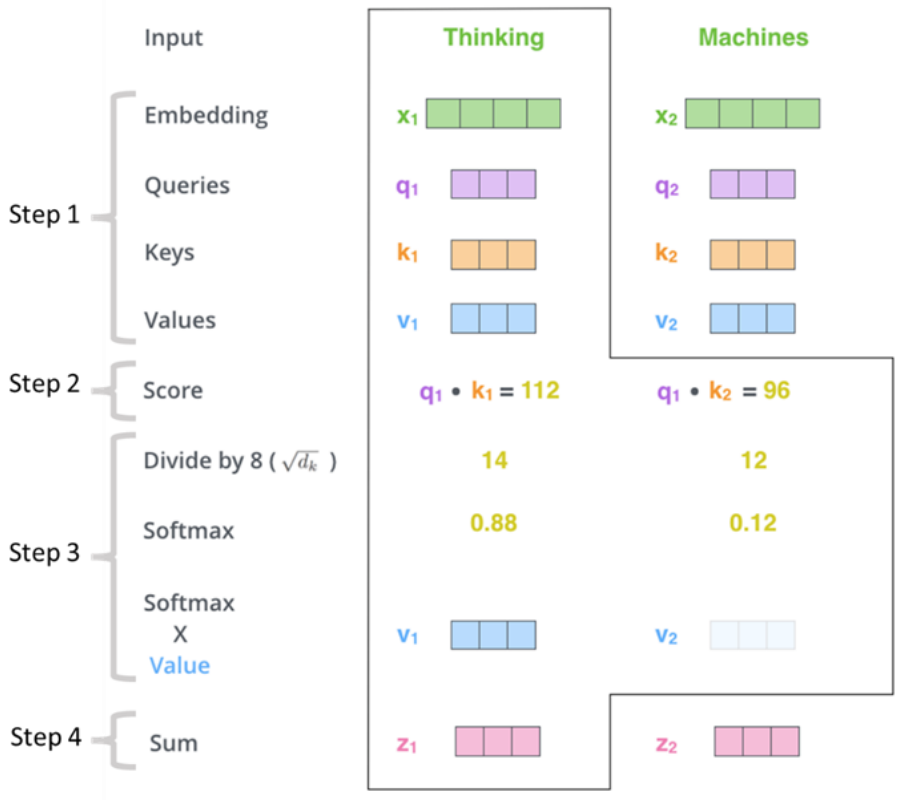
Appendix A: Self-attention mechanism



The four steps involved in the scaled dot-product attention function. Figure adapted from reference 1.

1. A query, key and value vector is created for each position in the input sequence $X_{1:n}$ by multiplying each word embedding by a query, key and value matrices. [1,2]

   Note: The query, key and value matrices are trained during pre-training of the model. [1,2]

2. For each position in the input sequence $X_{1:n}$, a set of scores is created by computing the dot product between its query vector and all query vectors in the sequence. These scores are scaled before application of the softmax function. [1,2]

   Note: In the original paper, scores are scaled by the square root of the length of the key vector.[1,2]

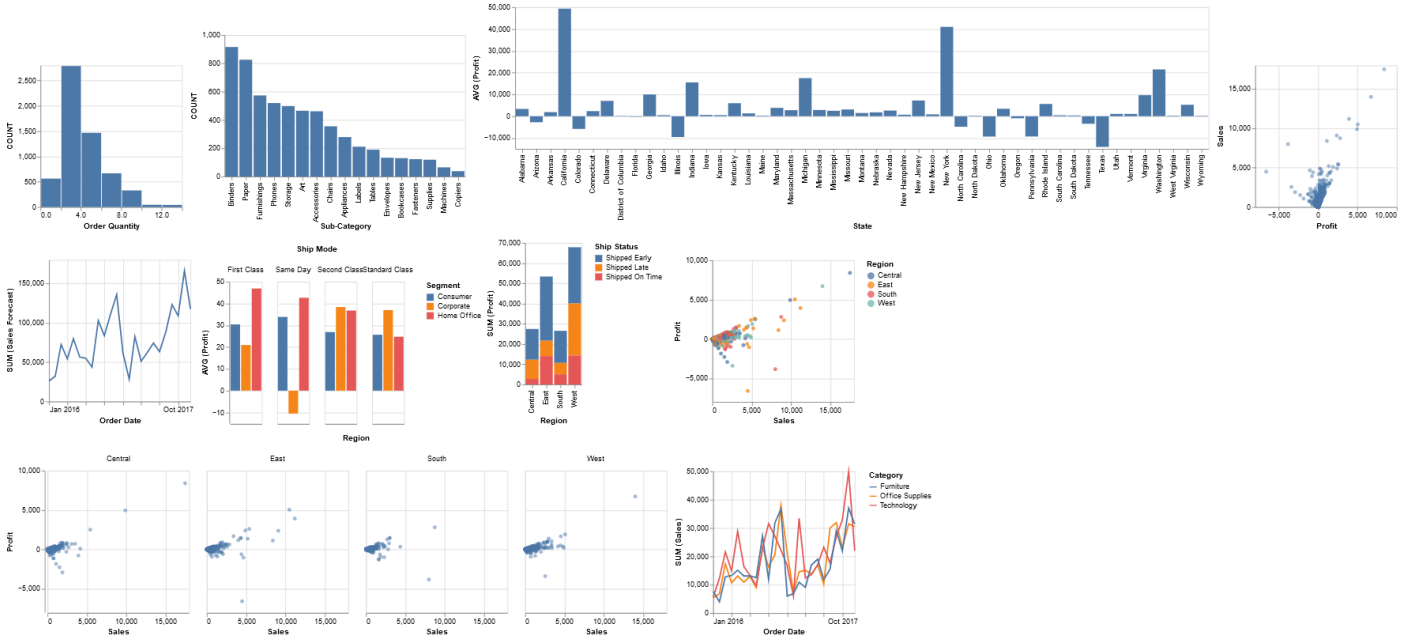
3. For each position in the input sequence $X_{1:n}$, a set of weighted value vectors is created by multiplying the set of softmax scores by its value vector. [1,2]

4. The sets of weighted value vectors are summed for each position in the input sequence $X_{1:n}$. The resulting vectors in each position are the contextual representations of the respective word embeddings in the input sequence. [1,2]

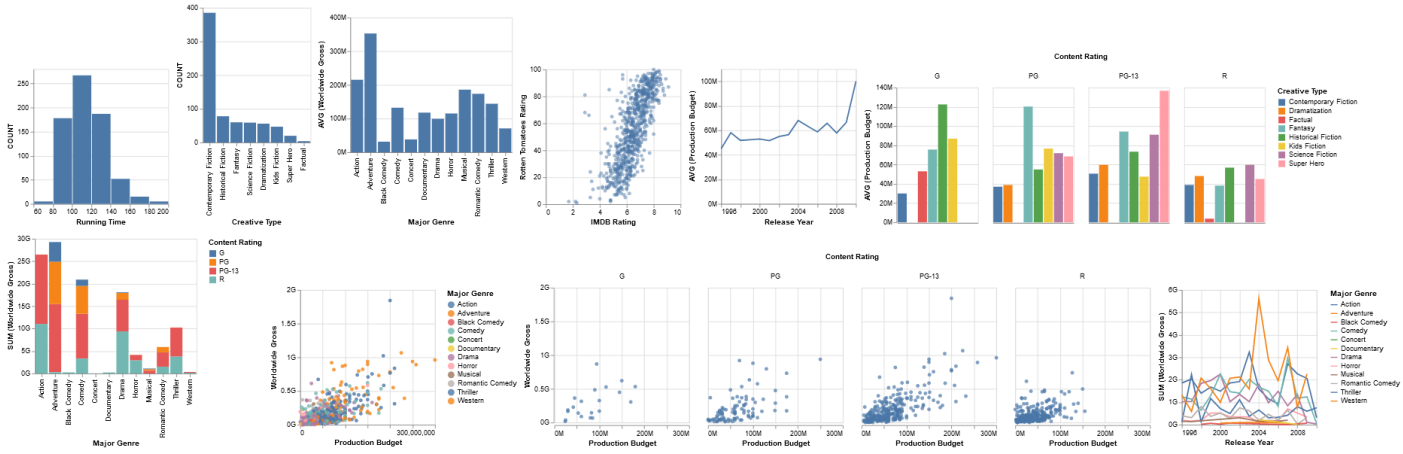
Appendix B: NLV corpus visualisations

The below visualisations are those used by Arjun Srinivasan and co-workers when collecting the NLV corpus.[3] The corpus provided 755 NL commands associated with these visualisations.

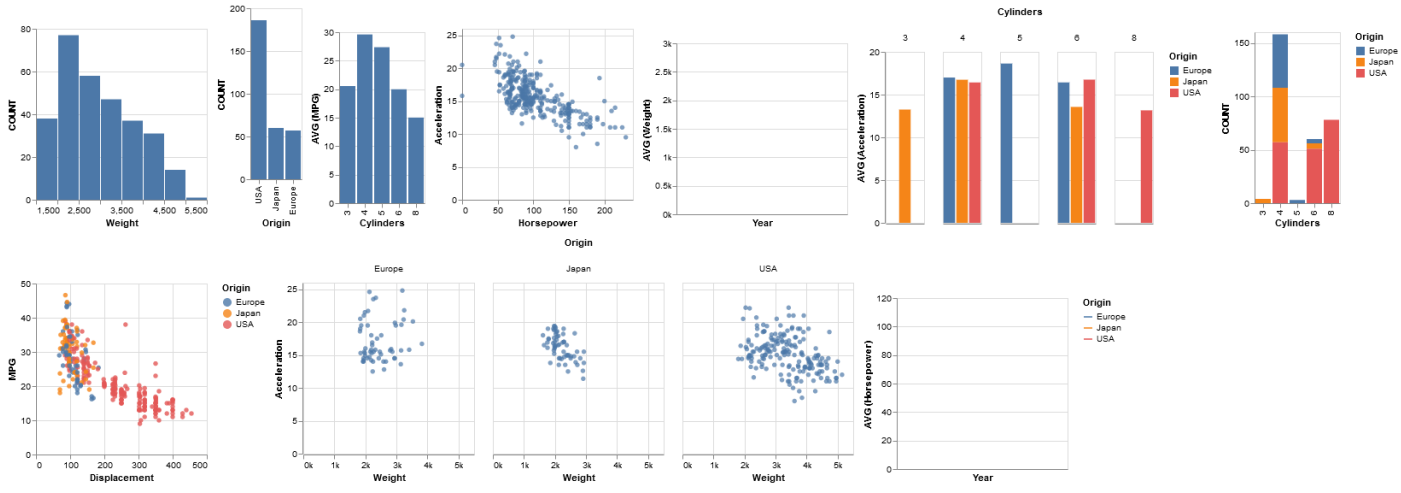
Superstore data



Movies data



Cars data



Appendix C: Research objective one, two and three - Data sources

The below datasets underpin all (NL command, Vega-Lite specification) pairs used during research objectives one, two and three.

Dataset	Train		Test	
	n	%	n	%
Cars	255	39.8	57	35.4
Euro	10	1.6	2	1.2
Movies	209	32.7	57	35.4
Superstore	166	25.9	45	28.0
Total	640	100.0	161	100.0

Table 1. Proportional breakdown of the data sources which underpin the train and test set. Where, n is the number of NL prompts, % is the % of all prompts and M is the median length in number of words.

Superstore data

Product-level sales information from order date to profit. The dataset contains 5,899 records and was used during collection of the NLV corpus and testing of NL4DV.

Variable	Level of measurement
Days to Ship	Ratio
Sales Forecast	Ratio
Ship Status	Nominal
Category	Nominal
City	Nominal
Country	Nominal
Customer Name	Nominal
Order Date	Ordinal
Order ID	Nominal
Product Name	Nominal
Profit	Interval
Quantity	Ratio
Region	Nominal
Sales	Ratio
Segment	Nominal
Ship Mode	Nominal
State	Nominal
Sub-Category	Nominal

Table 2. Features used in the raw Superstore dataset (https://github.com/nlvcorpus/nlvcorpus.github.io/blob/main/datasets/superstore.csv).

Movies data

Movie-level characteristics from worldwide gross to IMBD ratings. The dataset contains 709 records and was used during collection of the NLV corpus and testing of NL4DV.

Variable	Type
Title	Nominal
Worldwide Gross	Ratio
Production Budget	Ratio
Release Year	Ordinal
Content Rating	Nominal
Running Time	Ratio
Major Genre	Nominal
Creative Type	Nominal
Rotten Tomatoes Rating	Interval
IMDB Rating	Interval

Table 3. Features used in the raw Movies dataset (https://github.com/nlvcorpus/nlvcorpus.github.io/blob/main/datasets/movies.csv).

Cars data

Car model-level characteristics from year of release to weight. The dataset contains 303 records and was used during collection of the NLV corpus and testing of NL4DV.

Variable	Type

Model	Nominal
MPG (Miles per gallon)	Ratio
Cylinders	Ratio
Displacement	Ratio
Horsepower	Ratio
Weight	Ratio
Acceleration	Ratio
Year	Ordinal
Origin	Nominal

Table 4. Features used in the raw Cars dataset (https://github.com/nlvcorpus/nlvcorpus.github.io/blob/main/datasets/cars.csv).

European soccer players data

Player-level characteristics from age to goals scored. The dataset contains 552 records and was used during testing of NL4DV.

Variable	**Type**
Foot	Nominal
Name	Nominal
Position	Nominal
Club	Nominal
Country	Nominal
Age	Ratio
Salary	Ratio
Goals	Ratio

Table 5. Features used in the raw Euros dataset (https://github.com/nl4dv/nl4dv/blob/master/examples/assets/data/euro.csv)

Appendix D: Research objective four - Survey

Default Question Block

Natural language prompts for visualisations
Principal researcher: Dr Billy Pitchford
Principal investigator: Dr Pranava Madhyastha
Version 1.0 (15/11/2022)

We would like to invite you to take part in a research study. Before you decide whether you would like to take part it is important that you understand why the research is being done and what it would involve for you. Please take time to read the following information carefully. If you have any questions or concerns regarding the study, please contact Dr Billy Pitchford (billy.pitchford@city.ac.uk).

What is the purpose of the study?
The goal of this survey is collect natural language commands related to a range of data visualisations. These will be used as part of a research project investigating prospective natural language interfaces for data visualisation tools. The study is being completed as part of the requirements for a MSc Data Science degree at City, University of London and is funded via a level 7 degree apprenticeship.

Why have I been invited to take part?
This study is currently only open to English language speakers. Please note the study is anonymous. If you are a student or employee at City University, choosing to take part or not take part will have no bearing on future activities at City.

Do I have to take part?
Participation is voluntary, and you can choose not to particpate in part or all of the survey. Furthermore, you can withdraw at any stage without giving a reason. Please note only data from submitted surveys will be retained for further use.

What do I have to do if I take part?
The study will take approximately 7-10 minutes to complete. You will be asked to propose natural language commands for a set of 13 data visualisations. These visualisations have been created using publicly available Higher Education datasets.

What will happen to the results?
The data will contribute to a Masters thesis. If the study is successful, the data collected may contribute to a research article and be made available as part of an open-source repository. Please note no personally identifiable information will be collected.

What are the possible disadvantages and risks of taking part?
This survey is anonymous and the risks involved are no greater than those involved during routine web surveys.

What are the possible benefits of taking part?
An indirect benefit is contributing to an active research area motivated by the benefits of lowering the barrier of entry to data analytics. The research being conducted is novel and consequently a research article could be published if it's successful, thereby adding to the body of scientific knowledge in this area.

Expenses and Payments
You will not be financially compensated for your participation.

What should I do if I want to take part?
If you would like to take part, please provide consent by ticking the relevant tick boxes below.

What if there is a problem?

If you have any problems, concerns or questions about this study, you should ask to speak to a member of the research team. If you remain unhappy and wish to complain formally, you can do this through City's complaints procedure. To complain about the study, you need to phone 020 7040 3040. You can then ask to speak to the Secretary to Senate Research Ethics Committee and inform them that the name of the project is Natural language prompts for visualisations.

Conflicts of interests

We have no conflicts of interest to disclose.

Who has reviewed the study?

This study has been approved by City, University of London Computer Science Research Ethics Committee.

Thank you for taking the time to read this information sheet.

Consent form

I confirm that I have read and understood the participant information above (Version 1.0, 15/11/2022) and any questions I have asked have been answered satisfactorily.

- ○ Yes
- ○ No

I understand that my participation is voluntary and that I am free to withdraw without giving a reason without being penalised or disadvantaged.

- ○ Yes
- ○ No

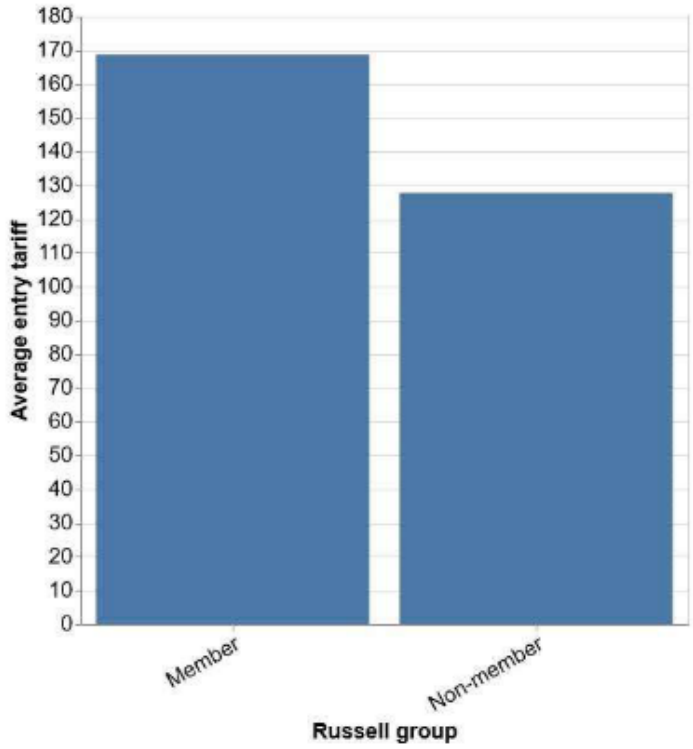
I understand that the anonymous data collected may be made open access to underpin journal publication.

- ○ Yes
- ○ No

I agree to take part in this study.

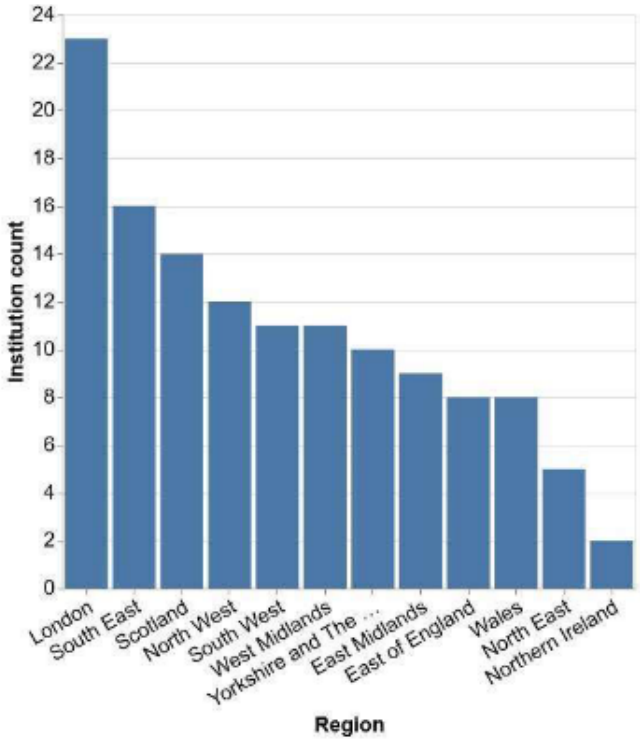
- ○ Yes
- ○ No

Institution	Russell group	Entry tariff	...
Anglia Ruskin	Non-member	109	...
Bath Spa	Non-member	124	...
Bournemouth	Non-member	116	...
Brighton	Non-member	117	...
Brunel	Non-member	123	...
...	...	...	...



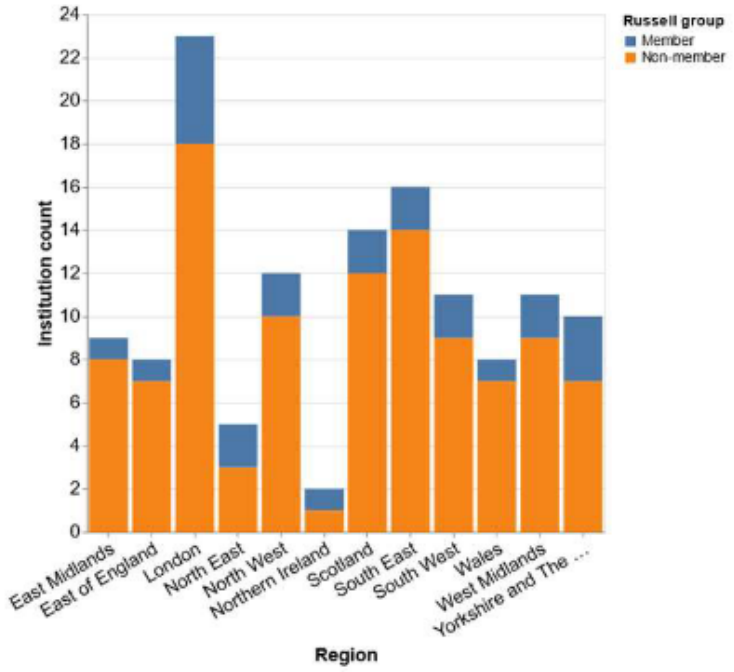
Propose one or more natural language statements you would enter into a visualisation system to generate the graph above from the dataset shown.

Institution	Russell group	Region	...
Anglia Ruskin	Non-member	East of England	...
Bath Spa	Non-member	South West	...
Bournemouth	Non-member	South West	...
Brighton	Non-member	South East	...
Brunel	Non-member	London	...
...	...	...	...



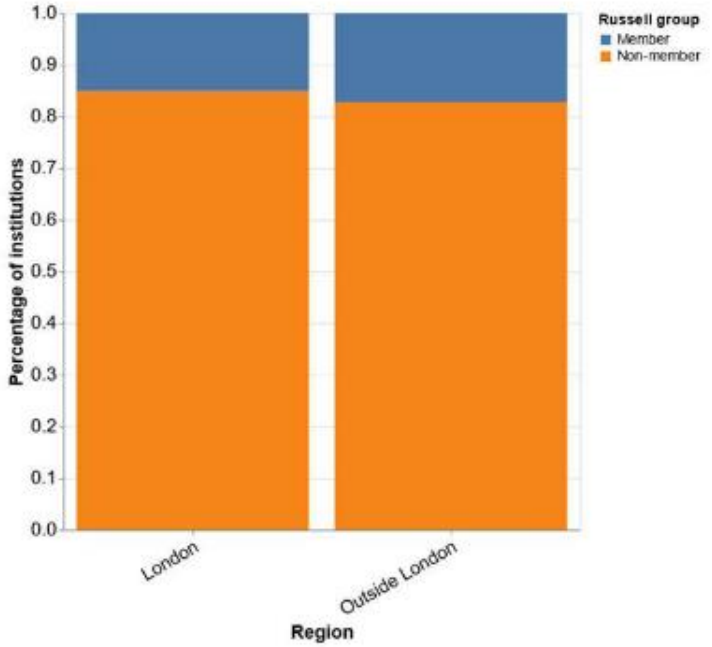
Propose one or more natural language statements you would enter into a visualisation system to generate the graph above from the dataset shown.

Institution	Russell group	Region	...
Anglia Ruskin	Non-member	East of England	...
Bath Spa	Non-member	South West	...
Bournemouth	Non-member	South West	...
Brighton	Non-member	South East	...
Brunel	Non-member	London	...
...	...	...	...



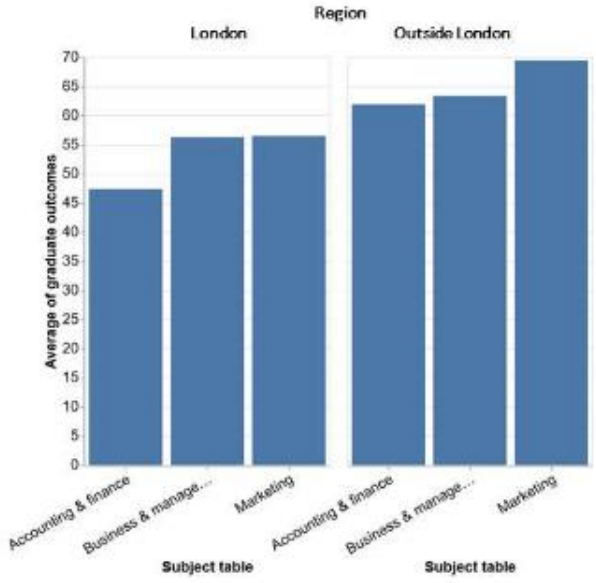
Propose one or more natural language statements you would enter into a visualisation system to generate the graph above from the dataset shown.

Institution	Russell group	Region	...
Anglia Ruskin	Non-member	East of England	...
Bath Spa	Non-member	South West	...
Bournemouth	Non-member	South West	...
Brighton	Non-member	South East	...
Brunel	Non-member	London	...
...	...	...	...



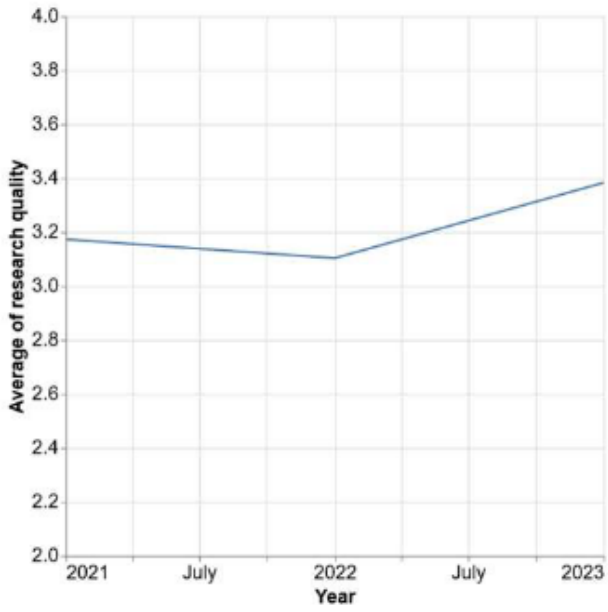
Propose one or more natural language statements you would enter into a visualisation system to generate the graph above from the dataset shown.

Subject table	Institution	Region	Graduate outcomes	...
Accounting & finance	De Montfort	Outside London	60	...
Business & management	De Montfort	Outside London	56	...
Marketing	De Montfort	Outside London	60	...
Accounting & finance	Loughborough	Outside London	78	...
Business & management	Loughborough	Outside London	88	...
...	...	...	...	...



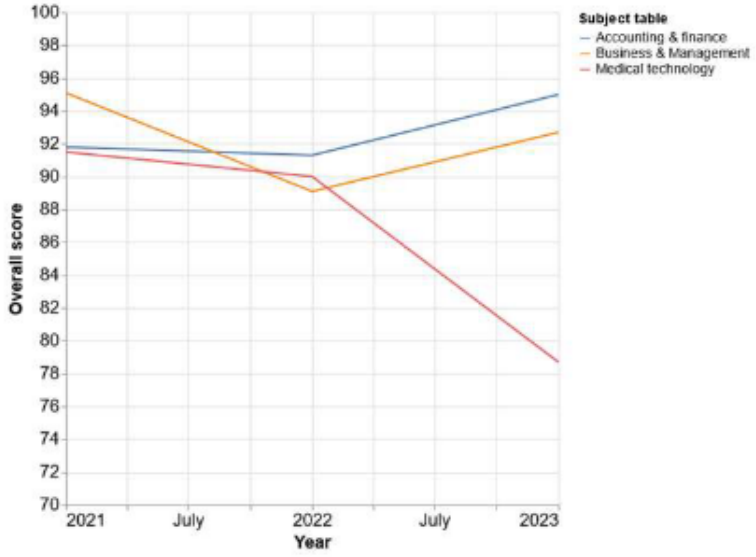
Propose one or more natural language statements you would enter into a visualisation system to generate the graph above from the dataset shown.

Year	Subject table	Research quality	...
2023	Accounting & finance	3.51	...
2022	Accounting & finance	3.23	...
2021	Accounting & finance	3.23	...
2023	Business & management	4.14	...
2022	Business & management	3.69	...
2021	Business & management	3.85	...
...	...	...	...

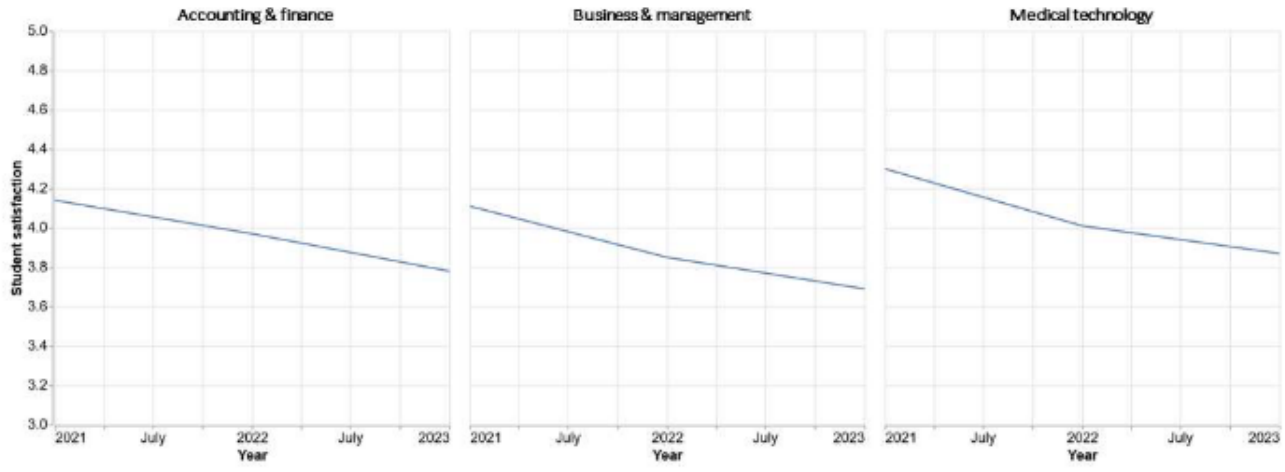


Propose one or more natural language statements you would enter into a visualisation system to generate the graph above from the dataset shown.

Year	Subject table	Overall score	...
2023	Accounting & finance	95.0	...
2022	Accounting & finance	91.3	...
2021	Accounting & finance	91.8	...
2023	Business & management	92.7	...
2022	Business & management	89.1	...
2021	Business & management	95.1	...
...	...	...	...



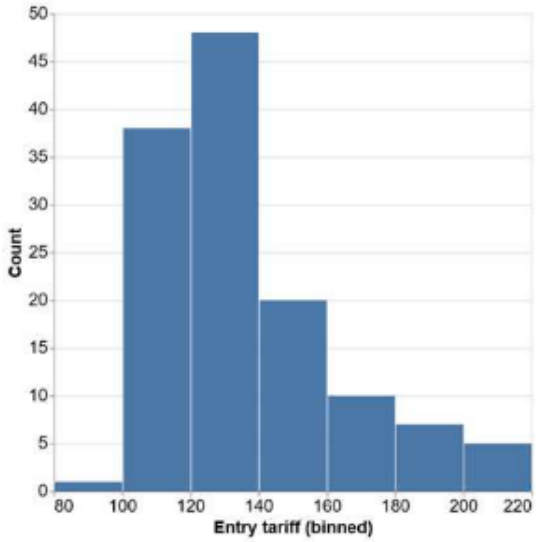
Propose one or more natural language statements you would enter into a visualisation system to generate the graph above from the dataset shown.



Year	Subject table	Student satisfaction	...
2023	Accounting & finance	3.78	...
2022	Accounting & finance	3.97	...
2021	Accounting & finance	4.14	...
2023	Business & management	3.69	...
2022	Business & management	3.85	...
2021	Business & management	4.11	...
...	...	...	...

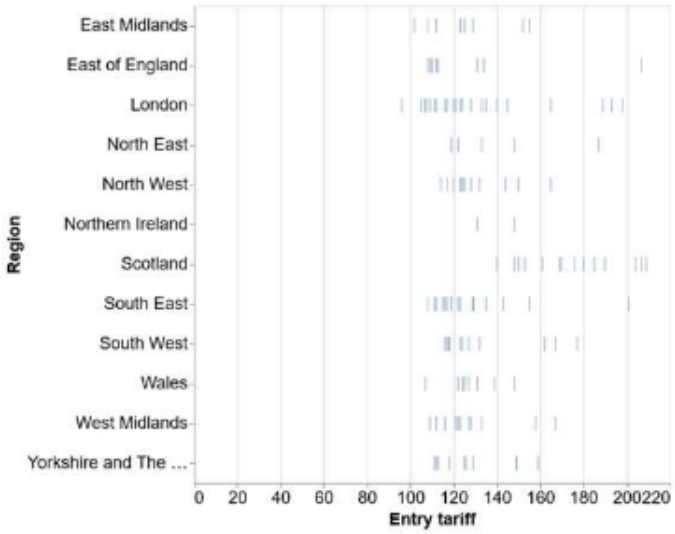
Propose one or more natural language statements you would enter into a visualisation system to generate the graph above from the dataset shown.

Institution	Russell group	Entry tariff	...
Anglia Ruskin	Non-member	109	...
Bath Spa	Non-member	124	...
Bournemouth	Non-member	116	...
Brighton	Non-member	117	...
Brunel	Non-member	123	...
...	...	...	...



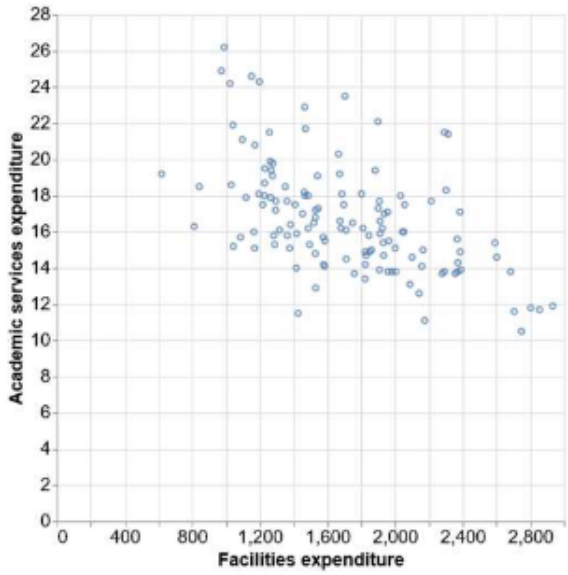
Propose one or more natural language statements you would enter into a visualisation system to generate the graph above from the dataset shown.

Institution	Region	Entrytariff	...
Anglia Ruskin	East of England	109	...
Bath Spa	South West	124	...
Bournemouth	South West	116	...
Brighton	South East	117	...
Brunel	London	123	...
...	...	...	...



Propose one or more natural language statements you would enter into a visualisation system to generate the graph above from the dataset shown.

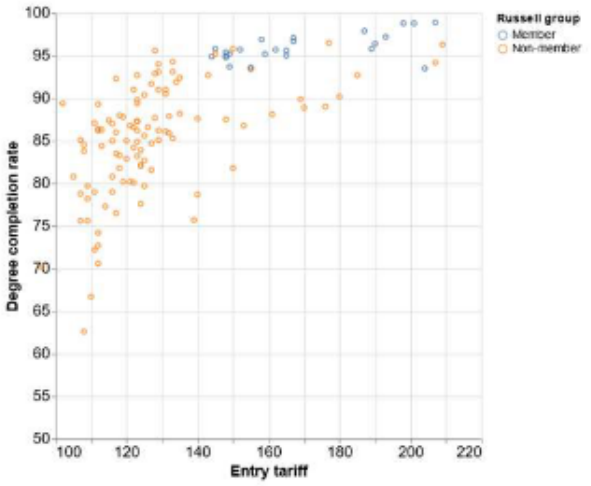
Institution	Academic services expenditure	Facilities expenditure	...
Anglia Ruskin	19.4	1267	...
Bath Spa	16.6	1674	...
Bournemouth	19.2	1672	...
Brighton	19.1	1801	...
Brunel	17.9	1118	...
...	...	...	...



Propose one or more natural language statements you would enter into a visualisation

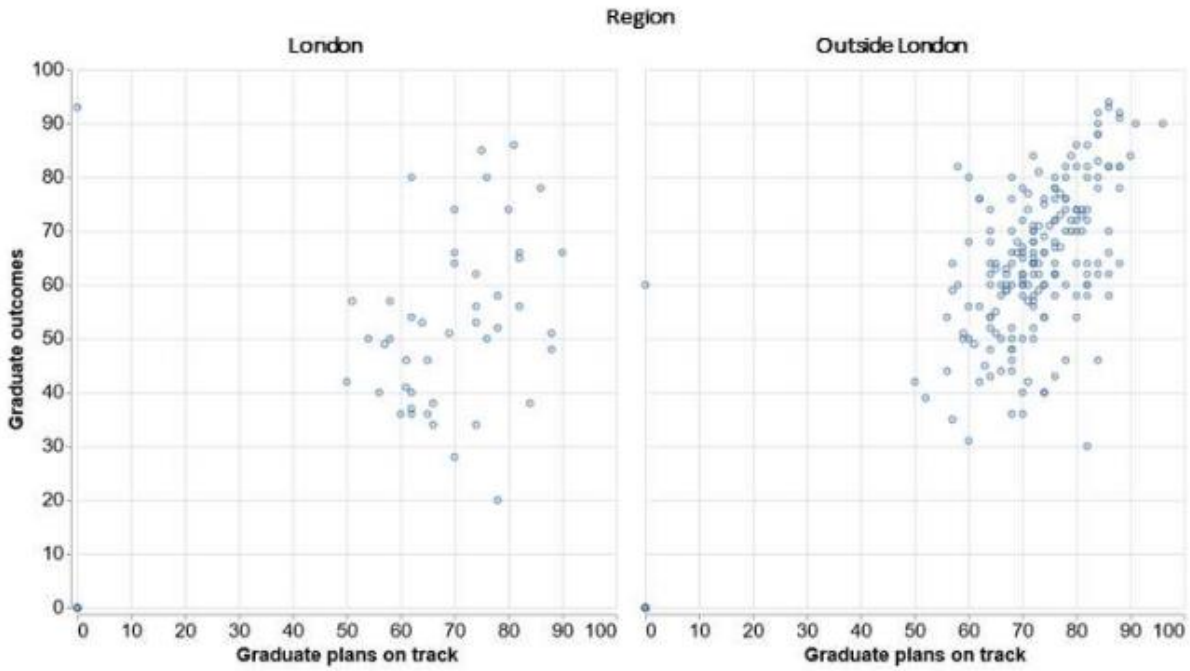
system to generate the graph above from the dataset shown.

Institution	Russell group	Degree completion rate	Entry tariff	...
Anglia Ruskin	Non-member	78.2	109	...
Bath Spa	Non-member	82.1	124	...
Bournemouth	Non-member	87.0	116	...
Brighton	Non-member	86.0	117	...
Brunel	Non-member	89.4	124	...
...	...	...	...	...



Propose one or more natural language statements you would enter into a visualisation system to generate the graph above from the dataset shown.

Block 1



Subject table	Institution	Region	Graduate outcomes	Graduate plans on track	...
Accounting & finance	De Montfort	Outside London	60.0	74.0	...
Business & management	De Montfort	Outside London	56.0	60.0	...
Marketing	De Montfort	Outside London	60.0	58.0	
Accounting & finance	Loughborough	Outside London	78.0	84.0	...
Business & management	Loughborough	Outside London	88.0	84.0	...
...	...	...	...	...	...

Propose one or more natural language statements you would enter into a visualisation system to generate the graph above from the dataset shown.

Appendix E

Research Ethics Review Form: BSc, MSc and MA Projects

Computer Science Research Ethics Committee (CSREC)

http://www.city.ac.uk/department-computer-science/research-ethics

Undergraduate and postgraduate students undertaking their final project in the Department of Computer Science are required to consider the ethics of their project work and to ensure that it complies with research ethics guidelines. In some cases, a project will need approval from an ethics committee before it can proceed. Usually, but not always, this will be because the student is involving other people ("participants") in the project.

In order to ensure that appropriate consideration is given to ethical issues, all students must complete this form and attach it to their project proposal document. There are two parts:

PART A: Ethics Checklist. All students must complete this part. The checklist identifies whether the project requires ethical approval and, if so, where to apply for approval.

PART B: Ethics Proportionate Review Form. Students who have answered "no" to all questions in A1, A2 and A3 and "yes" to question 4 in A4 in the ethics checklist must complete this part. The project supervisor has delegated authority to provide approval in such cases that are considered to involve MINIMAL risk.
The approval may be ***provisional*** – *identifying the planned research as* likely to involve MINIMAL RISK. In such cases you must additionally seek ***full approval*** from the supervisor as the project progresses and details are established. ***Full approval*** must be acquired in writing, before beginning the planned research.

A.1 If you answer YES to any of the questions in this block, you must apply to an appropriate external ethics committee for approval and log this approval as an External Application through Research Ethics Online - https://ethics.city.ac.uk/		*Delete as appropriate*
1.1	Does your research require approval from the National Research Ethics Service (NRES)? *e.g. because you are recruiting current NHS patients or staff?* *If you are unsure try - https://www.hra.nhs.uk/approvals-amendments/what-approvals-do-i-need/*	**NO**
1.2	Will you recruit participants who fall under the auspices of the Mental Capacity Act? *Such research needs to be approved by an external ethics committee such as NRES or the Social Care Research Ethics Committee - http://www.scie.org.uk/research/ethics-committee/*	**NO**
1.3	Will you recruit any participants who are currently under the auspices of the Criminal Justice System, for example, but not limited to, people on remand, prisoners and those on probation? *Such research needs to be authorised by the ethics approval system of the National Offender Management Service.*	**NO**
A.2 If you answer YES to any of the questions in this block, then unless you are applying to an external ethics committee, you must apply for approval from the Senate Research Ethics Committee (SREC) through Research Ethics Online - https://ethics.city.ac.uk/		*Delete as appropriate*

2.1	Does your research involve participants who are unable to give informed consent? *For example, but not limited to, people who may have a degree of learning disability or mental health problem, that means they are unable to make an informed decision on their own behalf.*	**NO**
2.2	Is there a risk that your research might lead to disclosures from participants concerning their involvement in illegal activities?	**NO**
2.3	Is there a risk that obscene and or illegal material may need to be accessed for your research study (including online content and other material)?	**NO**
2.4	Does your project involve participants disclosing information about special category or sensitive subjects? *For example, but not limited to: racial or ethnic origin; political opinions; religious beliefs; trade union membership; physical or mental health; sexual life; criminal offences and proceedings*	**NO**
2.5	Does your research involve you travelling to another country outside of the UK, where the Foreign & Commonwealth Office has issued a travel warning that affects the area in which you will study? *Please check the latest guidance from the FCO - http://www.fco.gov.uk/en/*	**NO**
2.6	Does your research involve invasive or intrusive procedures? *These may include, but are not limited to, electrical stimulation, heat, cold or bruising.*	**NO**
2.7	Does your research involve animals?	**NO**
2.8	Does your research involve the administration of drugs, placebos or other substances to study participants?	**NO**
A.3 If you answer YES to any of the questions in this block, then unless you are applying to an external ethics committee or the SREC, you must apply for approval from the Computer Science Research Ethics Committee (CSREC) through Research Ethics Online - https://ethics.city.ac.uk/ **Depending on the level of risk associated with your application, it may be referred to the Senate Research Ethics Committee.**		*Delete as appropriate*
3.1	Does your research involve participants who are under the age of 18?	**NO**
3.2	Does your research involve adults who are vulnerable because of their social, psychological or medical circumstances (vulnerable adults)? *This includes adults with cognitive and / or learning disabilities, adults with physical disabilities and older people.*	**NO**
3.3	Are participants recruited because they are staff or students of City, University of London? *For example, students studying on a particular course or module.* *If yes, then approval is also required from the Head of Department or Programme Director.*	**NO**
3.4	Does your research involve intentional deception of participants?	**NO**

3.5	Does your research involve participants taking part without their informed consent?	**NO**
3.5	Is the risk posed to participants greater than that in normal working life?	**NO**
3.7	Is the risk posed to you, the researcher(s), greater than that in normal working life?	**NO**
A.4 If you answer YES to the following question and your answers to all other questions in sections A1, A2 and A3 are NO, then your project is deemed to be of MINIMAL RISK. **If this is the case, then you can apply for approval through your supervisor under PROPORTIONATE REVIEW. You do so by completing PART B of this form.** **If you have answered NO to all questions on this form, then your project does not require ethical approval. You should submit and retain this form as evidence of this.**		*Delete as appropriate*
4	Does your project involve human participants or their identifiable personal data? *For example, as interviewees, respondents to a survey or participants in testing.*	**Yes**

PART B: Ethics Proportionate Review Form

If you answered YES to question 4 and NO to all other questions in sections A1, A2 and A3 in PART A of this form, then you may use PART B of this form to submit an application for a proportionate ethics review of your project. Your project supervisor has delegated authority to review and approve this application under proportionate review. You must receive final approval from your supervisor in writing before beginning the planned research.

However, if you cannot provide all the required attachments (see B.3) with your project proposal (e.g. because you have not yet written the consent forms, interview schedules etc), the approval from your supervisor will be ***provisional***. You **must** submit the missing items to your supervisor for approval prior to commencing these parts of your project. Once again, you must receive written confirmation from your supervisor that any provisional approval has been superseded by with ***full approval*** of the planned activity as detailed in the full documents. **Failure to follow this procedure and demonstrate that final approval has been achieved may result in you failing the project module.**

Your supervisor may ask you to submit a full ethics application through Research Ethics Online, for instance if they are unable to approve your application, if the level of risks associated with your project change, or if you need an approval letter from the CSREC for an external organisation.

B.1 The following questions must be answered fully. All grey instructions must be removed.		*Delete as appropriate*
1.1.	Will you ensure that participants taking part in your project are fully informed about the purpose of the research?	Yes
1.2	Will you ensure that participants taking part in your project are fully informed about the procedures affecting them or affecting any information collected about them, including information about how the data will be used, to whom it will be disclosed, and how long it will be kept?	Yes
1.3	When people agree to participate in your project, will it be made clear to them that they may withdraw (i.e. not participate) at any time without any penalty?	Yes
1.4	Will consent be obtained from the participants in your project? Consent from participants will be necessary if you plan to involve them in your project or if you plan to use identifiable personal data from existing records. "Identifiable personal data" means data relating to a living person who might be identifiable if the record includes their name, username, student id, DNA, fingerprint, address, etc. *If YES, you must attach drafts of the participant information sheet(s) and consent form(s) that you will use in section B.3 or, in the case of an existing dataset, provide details of how consent has been obtained.* *You must also retain the completed forms for subsequent inspection. Failure to provide the completed consent request forms will result in withdrawal of any earlier ethical approval of your project.*	Yes
1.5	Have you made arrangements to ensure that material and/or private information obtained from or about the participating individuals will remain confidential?	Yes

B.2 If the answer to the following question (B2) is YES, you must provide details		*Delete as appropriate*
2	Will the research be conducted in the participant's home or other non-University location? *If **YES**, you must provide details of how your safety will be ensured.*	No

B.3 Attachments **ALL of the following documents MUST be provided to supervisors if applicable.** **All must be considered prior to final approval by supervisors.** **A written record of final approval must be provided and retained.**	***YES***	***NO***	***Not Applicable***
Details on how safety will be assured in any non-University location, including risk assessment if required (see B2)			NA
Details of arrangements to ensure that material and/or private information obtained from or about the participating individuals will remain confidential (see B1.5) *Any personal data must be acquired, stored and made accessible in ways that are GDPR compliant.*			NA
Full protocol for any workshops or interviews**			NA
Participant information sheet(s)**	Yes		
Consent form(s)**	Yes		NA
Questionnaire(s)** *sharing a Qualtrics survey with your supervisor is recommended.*	Yes		NA
Topic guide(s) for interviews and focus groups**			NA
Permission from external organisations or Head of Department** *e.g. for recruitment of participants*			NA

***If these items are not available at the time of submitting your project proposal, then **provisional approval** can still be given, under the condition that you must submit the final versions of all items to your supervisor for approval at a later date. **All** such items **must** be seen and approved by your supervisor before the activity for which they are needed begins. Written evidence of **final approval** of your planned activity must be acquired from your supervisor before you commence.*

Changes

If your plans change and any aspects of your research that are documented in the approval process change as a consequence, then any approval acquired is invalid. If issues addressed in Part A (the checklist) are affected, then you must complete the approval process again and establish the kind of approval that is required. If issues addressed in Part B are affected, then you must forward updated documentation to your supervisor and have received written confirmation of approval of the revised activity before proceeding.

Templates for Consent and Information

You must use the templates provided by the University as the basis for your participant information sheets and consent forms. You **must** adapt them according to the needs of your project before you submit them for consideration.

Participant Information Sheets, Consent Forms and Protocols must be consistent. Please ensure that this is the case prior to seeking approval. Failure to do so will slow down the approval process.

We strongly recommend using Qualtrics to produce digital information sheets and consent forms.

Further Information

http://www.city.ac.uk/department-computer-science/research-ethics

https://www.city.ac.uk/research/ethics/how-to-apply/participant-recruitment

https://www.city.ac.uk/research/ethics

Appendix F: Ethics approval

Pitchford, William

From: Madhyastha, Pranava
Sent: 16 November 2022 16:25
To: Pitchford, William
Subject: Re: Project - HE dataset

Hi Billy,

All looks good! It is approved from my side (this email is the official proof of the approval, just in case it is needed later).

All the best,
Pranava

_____
From: Pitchford, William <Billy.Pitchford@city.ac.uk>
Sent: 16 November 2022 13:28
To: Madhyastha, Pranava
Subject: RE: Project - HE dataset

Hi Pranava,

No problem. Please find attached the completed form.

Please note I've put 'No' for Q3.3: 'Are participants recruited because they are staff or students of City, University of London?' I'll distribute the survey via Reddit instead. From looking at Part B of the ethics review form, this means you have delegated authority to review and approve the form and approval by CSREC would not be required.

I've now created the relevant survey, which also contains the participant information sheet and consent form. It can be accessed here: https://cityunilondon.eu.qualtrics.com/jfe/form/SV_3h3OGjD42B3lHh4

Please let me know if you require any further information.

Best wishes,

Billy

-----Original Message-----
From: Madhyastha, Pranava <Pranava.Madhyastha@city.ac.uk>
Sent: 14 November 2022 15:06
To: Pitchford, William <Billy.Pitchford@city.ac.uk>
Subject: Re: Project - HE dataset

That sounds good!

Would you fill the ethics form and send me over. I will then approve it, that way there is a record of it, in case the university needs it. I believe the previous form is empty.

Pranava

Appendix G: Research objective four - Data sources

The below datasets underpin all (NL command, Vega-Lite specification) pairs used during research objective four.

Dataset	Test	
	n	%
Main table	75	55.1
Sector performance at subject-level	31	22.8
City performance at subject-level	30	22.1
Total		

Table 6. Proportional breakdown of the data sources which underpin the Higher Education test set. Where, n is the number of NL prompts, % is the % of all prompts and M is the median length in number of words.

Main table

Institution-level results in the main table of the 2023 Complete University Guide. The dataset contains 128 records.

Variable	Type
institution	Nominal
Russell group	Nominal
region	Nominal
academic services expenditure	Ratio
degree completion rate	Ratio
entry tariff	Ratio
facilities expenditure	Ratio
good degree rate	Ratio
graduate prospects	Ratio
student satisfaction	Interval
research quality	Interval
research intensity	Ratio
student to staff ratio	Ratio
overall score	Ratio
overall rank	Ordinal

Table 7. Features in the raw dataset (https://github.com/earth1987/MSc-Thesis-Data/blob/main/2023_sector.csv).

Sector performance at subject-level

Institution-level results for three subject tables (Accounting & finance, Business & management and Marketing) in the 2023 Complete University Guide. The dataset contains 268 records.

Variable	Type
year	Ordinal
table	Nominal
institution	Nominal
Russell group	Nominal
region	Nominal
entry tariff	Ratio
student satisfaction	Interval
research quality	Interval
overall score	Ratio
rank	Ordinal
graduate outcomes	Ratio
graduate plans on track	Ratio

Table 8. Features in the raw dataset (https://github.com/earth1987/MSc-Thesis-Data/blob/main/Sector_sbj_by_yr.csv).

City performance at subject-level

Subject-level results for City in three subject tables (Accounting & finance, Business & management and Marketing) in the Complete University Guide from 2019 to 2023.

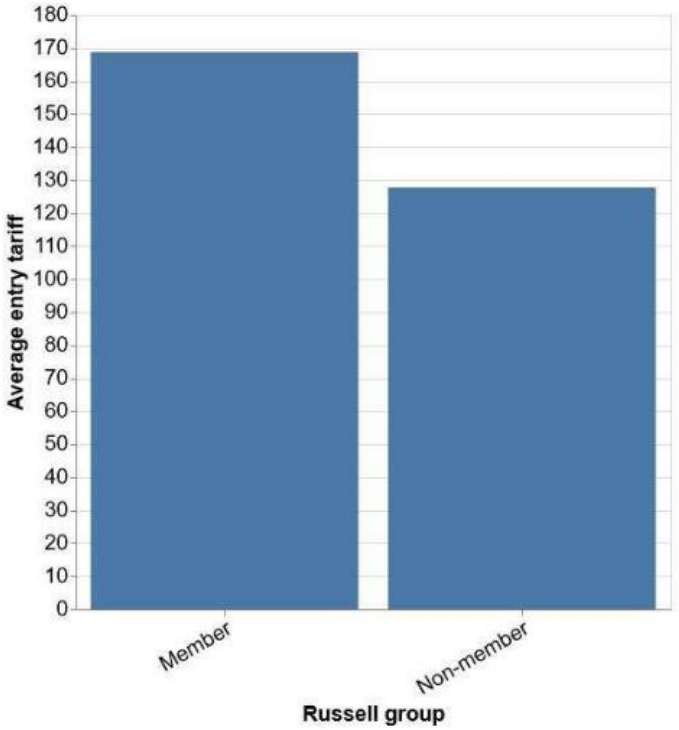
Variable	Type
year	Ordinal
table	Nominal
entry tariff	Ratio
student satisfaction	Interval
research quality	Interval
overall score	Ratio
rank	Ordinal

Table 9. Features in the raw dataset (https://github.com/earth1987/MSc-Thesis-Data/blob/main/Subject_by_yr.csv).

Appendix H: Research objective four - survey responses

Q1: Bar chart (basic)

Institution	Russell group	Entry tariff	...
Anglia Ruskin	Non-member	109	...
Bath Spa	Non-member	124	...
Bournemouth	Non-member	116	...
Brighton	Non-member	117	...
Brunel	Non-member	123	...
...	...	...	...



"Create a bar chart with Russell group member or non member on the x axis and tariff on the Y axis"

"what is the difference between the average entry tariff for members and non members of russell group universities"

"Average entry tariff of Russell group members"

"This graphs shows average entry tariff of Russell group members versus non-members"

"Plot a bar graph showing member vs non member entry tariffs"

"Show me the average entry tariff for a member of the Russell group"

"Create a bar chart showing the average entry tariff of Russell group members versus non-members"

"List member and non member institutions along the bottom axis of the graph"

"Differences in entry tariffs between Russell and non-Russell group universities"

"Create a bar chart showing the average entry tariffs of members and non members of Russel group universities"

"Cross tabulate entry tariff by Russell Group"

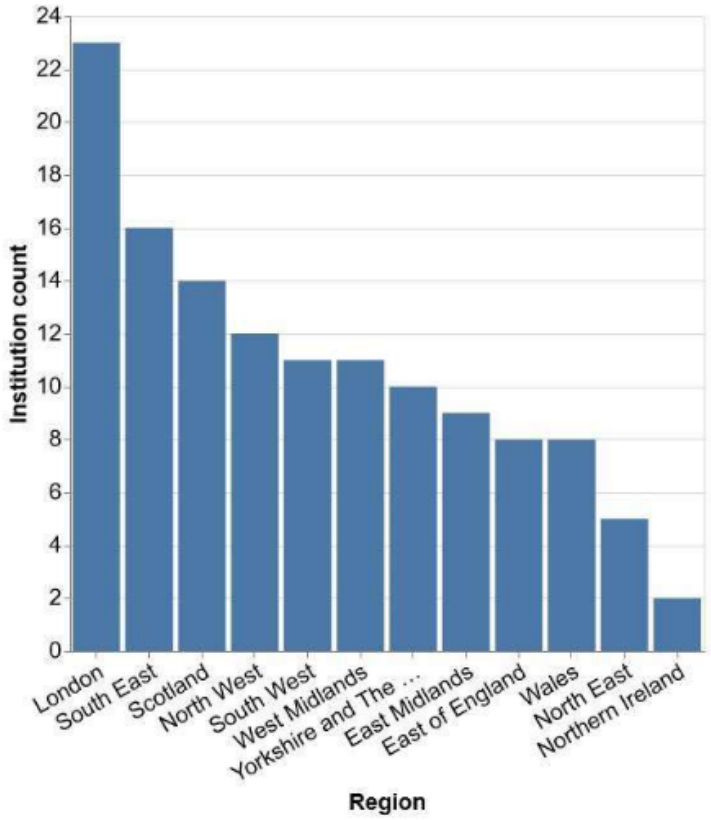
"create bar chart with russel group on the x axis and average entry tariff on the y axis"

"Average entry tariff value of member and non-member institutions of the Russel group"

"A graph showing the difference between member and non-member average entry tariffs"

Q2: Bar chart (sorted)

Institution	Russell group	Region	...
Anglia Ruskin	Non-member	East of England	...
Bath Spa	Non-member	South West	...
Bournemouth	Non-member	South West	...
Brighton	Non-member	South East	...
Brunel	Non-member	London	...
...	...	...	...



“Region column in x axis. Institution count in the Y axis”

“show me a graphic representation of the differences in the count of russell group universities between regions in the united kingdom”

“Institution count of the different regions across the uk”

“This graphic shows the Institution count across different Regions in the UK”

“Plot a bar graph showing institution count vs region”

“Show me the institution counts for the different regions of the UK”

“Create a bar graph showing the number of non-Russel group member institutions by region across the UK”

“Use the regions where institutions are based along the horizontal axis and the entry requirements along the vertical axis to generate a graph”

“Number of Institutions by region”

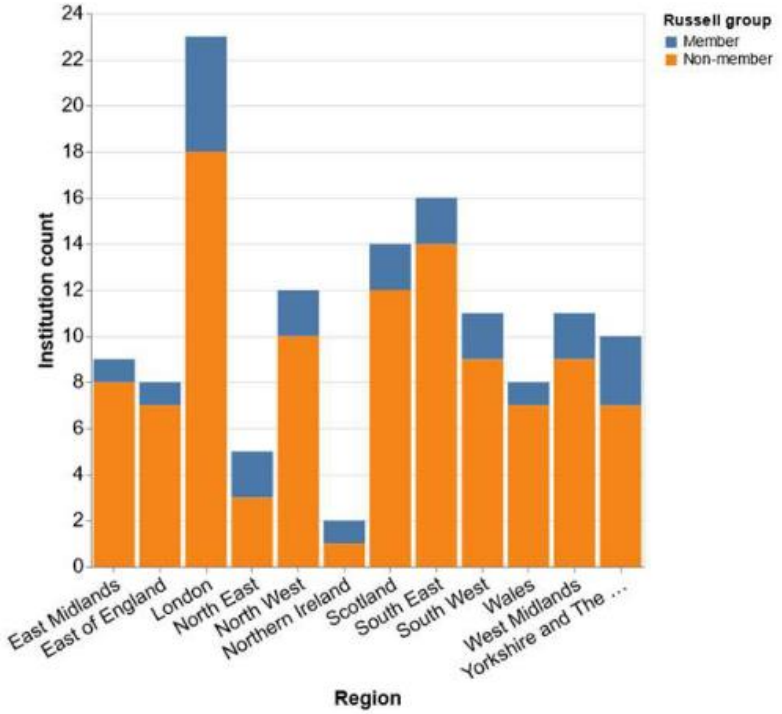
“Create a bar chart that shows how many university institutions there are (Russel Group or Non Russel Group members) in specified regions of the UK”

“Show me number of institutions by geographic region”

“create a bar chart with region on the x axis and institution count on the y axis”

Q3: Stacked bar chart

Institution	Russell group	Region	...
Anglia Ruskin	Non-member	East of England	...
Bath Spa	Non-member	South West	...
Bournemouth	Non-member	South West	...
Brighton	Non-member	South East	...
Brunel	Non-member	London	...
...	...	...	...



"Use region on x axis and q institution on y axis. Have orange bars for Russell group and blue of non Russel group"

"Show me a graphic representation of the institution count of universities between different regions in the UK, and separate Russell group members from non-members"

"Institution count across different regions in the uk including members and non members"

"This graphic shows the Institution count across different regions of the UK, split into Russell group members/non-members"

"Plot a bar graph showing institution count vs region split out into Russell group member vs non member"

"Plot a graph to show the institution counts for member and non-members of the Russell group in the different regions of the UK"

"Create a bar chart showing the number of institutions in different regions across the UK and show the number of Russell Group member and non-member institutions per region."

"Create a graph from the dataset depicting the number of Russell group and non-Russell group members in the different regions across the uk"

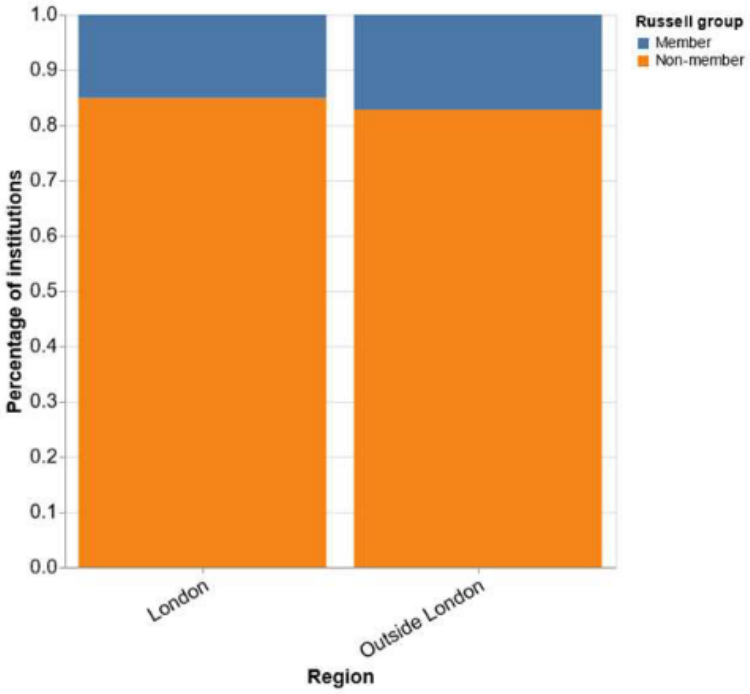
"Number of institutions by region and Russell group membership"

"Create a bar chart that shows how many university institutions there are in specified regions of the UK, and that uses coloured stacks to show how many are Russel Group members and how many are not for each region."

"Show me number of institutions by geographic region by type / Russell Group status"

Q4: Stacked bar chart (part-to-whole)

Institution	Russell group	Region	...
Anglia Ruskin	Non-member	East of England	...
Bath Spa	Non-member	South West	...
Bournemouth	Non-member	South West	...
Brighton	Non-member	South East	...
Brunel	Non-member	London	...
...	...	...	...



"Creat bar chart showing percentage of institutions inside London which are and are not Russell group and percentage of institutions outside London which are and are not Russell group"

"I want to see the difference in percentage of institutions that are members and non members of the Russell Group between London and Outside London"

"Percentage of institutions that are either members or non members within London and outside of London"

"This graphic shows the percentage of Institutions that are either members of non-members of the Russell group, this is shown for London and for Outside London as two separate groups"

"Plot a bar graph showing percentage of Russell group members in regions of London and Outside London"

"Show the percentage of institutions in London and outside of London in terms of how many members and non members of the Russell group"

"Create a bar chart showing the percentage of institutions in and outside of London and show the split of Russell Group member and non-members too"

"Create a graph showing the number of Russell group and non-Russell group institutions in the UK from both inside and outside London"

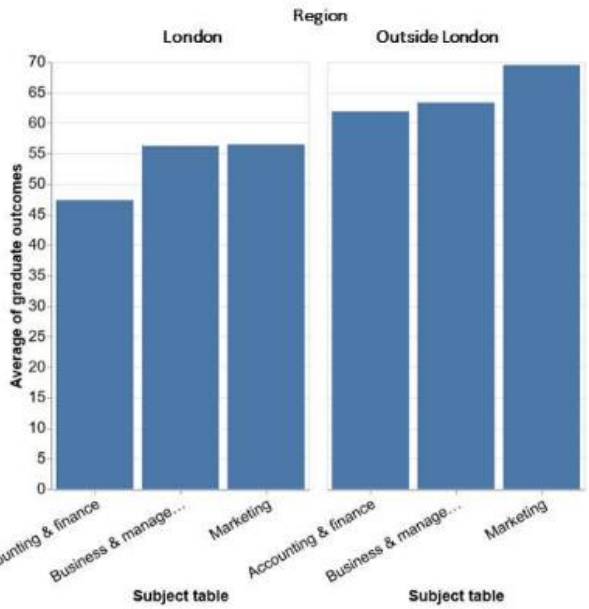
"Percentage of institutions in London compared to the outside of London, further divided by Russel group and non-Russel group"

"Create a bar chart showing the % of university institutions that are Russel Group and Non Russel Group members, comparing institutions in London with those Outside London"

"Show me number of institutions inside and outside of London by type / Russell Group status"

Q5: Bar chart (multiples)

Subject table	Institution	Region	Graduate outcomes	...
Accounting & finance	De Montfort	Outside London	60	...
Business & management	De Montfort	Outside London	56	...
Marketing	De Montfort	Outside London	60	...
Accounting & finance	Loughborough	Outside London	78	...
Business & management	Loughborough	Outside London	88	...
...	...	...	...	...



"On the x axis show subject on Y axis show average outcome. Order bars London and outside of london"

"Average of graduate outcomes within different subjects both in London and outside of London"

"This graphic shows the average graduate outcomes across different subjects studied. This is split into two separate groups, one for London and one for Outside London"

"Plot a bar graph showing average of graduate outcomes per subject, split into London and outside London regions"

"Plot a graph to show the average of graduate outcomes in London compared to outside London for the following subjects, accounting & finance, business & management, and marketing."

"Create a bar chart showing the average graduate outcomes per subject in and outside of London."

"Using the dataset shown, create a graph depicting the graduate outcomes for students across the different subjects from both inside and outside london."

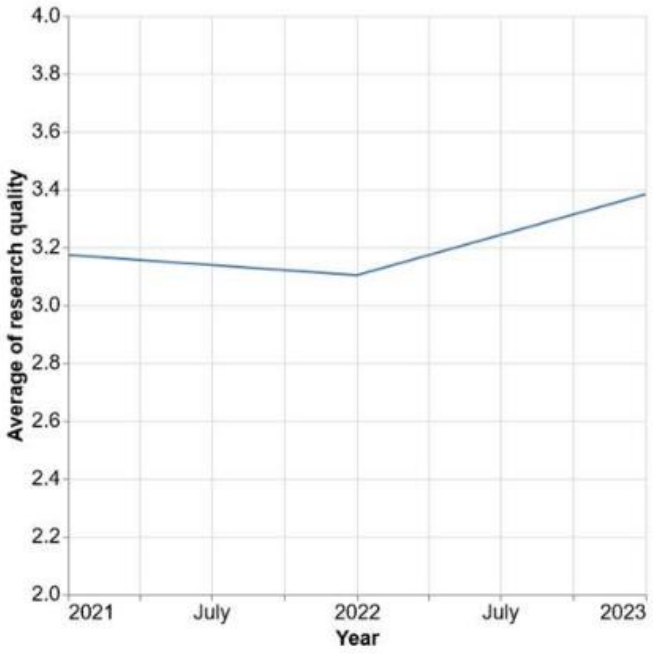
"Average graduate outcome by subject, comparing London and non-London institutions"

"Create a bar chart showing different course subjects at UK institutions and the average mark of graduates of those courses"

"Show me graduate outcomes for each subject by geographic region of university"

Q6: Line chart

Year	Subject table	Research quality	...
2023	Accounting & finance	3.51	...
2022	Accounting & finance	3.23	...
2021	Accounting & finance	3.23	...
2023	Business & management	4.14	...
2022	Business & management	3.69	...
2021	Business & management	3.85	...
...	...	...	...



"Create line graph with year on x axis with six month intervals. On y axis average research quality in intervals of 2 tenths"

"Show the average of research quality between the years of 2021 and 2023"

"This graph shows the average research quality from 2021 to 2023"

"Plot a line graph showing the average of research quality vs the year"

"Plot a graph to show the average of research quality between 2021 and 2023"

"Create a line graph showing the average research quality between 2021 and 2023"

"Using the information in the dataset shown, create a graph depicting the average research quality on a year-by-year basis"

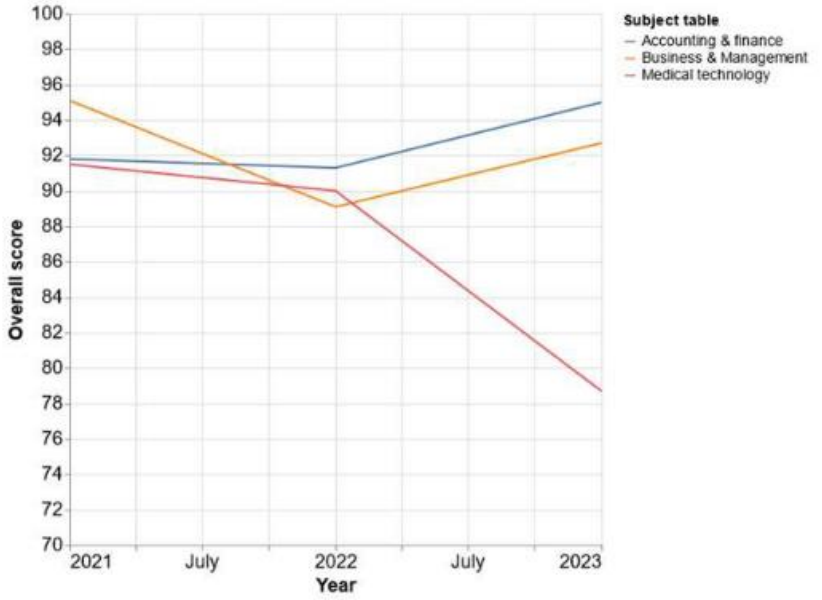
"Average research quality by year"

"Create a line graph showing how research quality changes each quarter across all courses at UK institutions"

"Show me research quality by year"

Q7: Line chart (coloured)

Year	Subject table	Overall score	...
2023	Accounting & finance	95.0	...
2022	Accounting & finance	91.3	...
2021	Accounting & finance	91.8	...
2023	Business & management	92.7	...
2022	Business & management	89.1	...
2021	Business & management	95.1	...
...	...	...	...



"Create a line graph with three different lines showing accounting, business and medical tech. Date in the x axis and score in the y axis"

"Plot the overall score of different subjects spanning the years from 2021 to 2023"

"This graph shows the overall score of three different subjects (Accounting & finance, Business & Management & Medical technology) from 2021 to 2023"

"Plot multiple line graphs showing overall score vs year with specific subjects"

"Show the overall score of each of accounting & finance, business and management, and medical technology between 2021 and 2023"

"Create a line graph showing the overall score per subject by year"

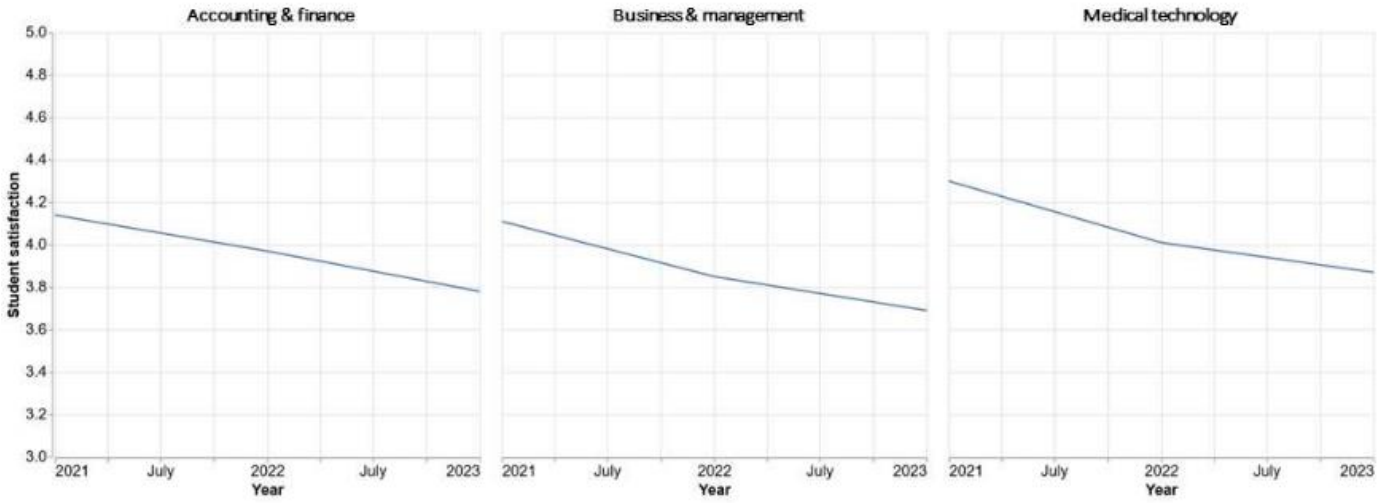
"Create a graph using the dataset shown, to depict a graph demonstrating the overall score of each subject matter on a year-by-year basis"

"Overall score by year across subjects"

"Create a line graph that plots the scores of three different subjects (Accounting & Finance, Business & Management, Medical Technology) and how overall scores have changed each quarter from 2021 to 2023"

"Show me score by year with respect to subject"

Q8: Line chart (multiples)



Year	Subject table	Student satisfaction	...
2023	Accounting & finance	3.78	...
2022	Accounting & finance	3.97	...
2021	Accounting & finance	4.14	...
2023	Business & management	3.69	...
2022	Business & management	3.85	...
2021	Business & management	4.11	...
...	...	...	...

"Create three deprecate line graphs. On one show accounting on another show business on the third show medical tech."

"Show the level of student satisfaction from Accounting and finance, business and management and medical technology across the years 2021 to 2023"

"The above graphs show student satisfaction from 2021 to 2023, this is separately shown for the following three subjects: Accounting & finance, Business & management, Medical technology"

"Plot multiple individual line graphs showing student satisfaction vs year for specific subjects"

"Plot three separate graphs to show student satisfaction between 2021 and 2023 for accounting & finance, business & management, and medical technology"

"Create three graphs but with the same y axis that shows the student satisfaction score by subject per year"

"Create individual graphs for each subject matter depicting student satisfaction by dear and institution"

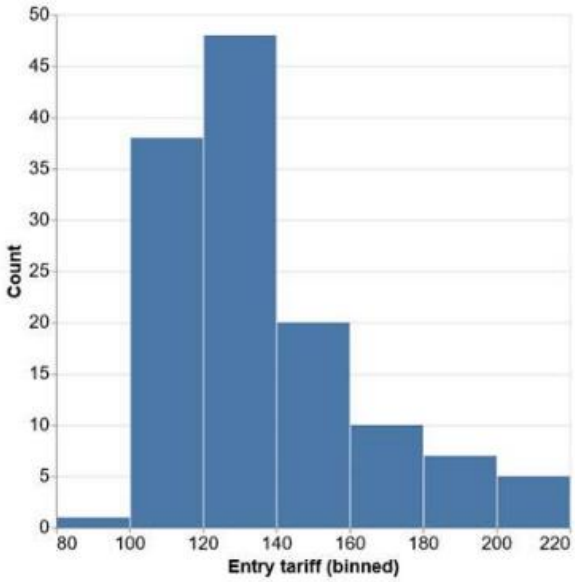
"Student satisfaction by year separated into separate line graphs for each subject"

"Create three separate line charts for each course that shows how student satisfaction has changed each quarter from 2021 to 2023."

"Show student satisfaction score by year by subject"

Q9: Histogram

Institution	Russell group	Entry tariff	...
Anglia Ruskin	Non-member	109	...
Bath Spa	Non-member	124	...
Bournemouth	Non-member	116	...
Brighton	Non-member	117	...
Brunel	Non-member	123	...
...	...	...	...



"Create a bar chart with entry tarif on the x axis in intervals of 20 on Y axis show number"

"Graph that shows the count on entrants against groups of Entry Tariffs"

"Plot a bar graph showing count vs entry tariff (binned)"

"Plot a graph to show the count of entry tariffs"

"Create a bar chart showing the entry tariff"

"Create a graphs depicting the entry requirements based on the number of institutions"

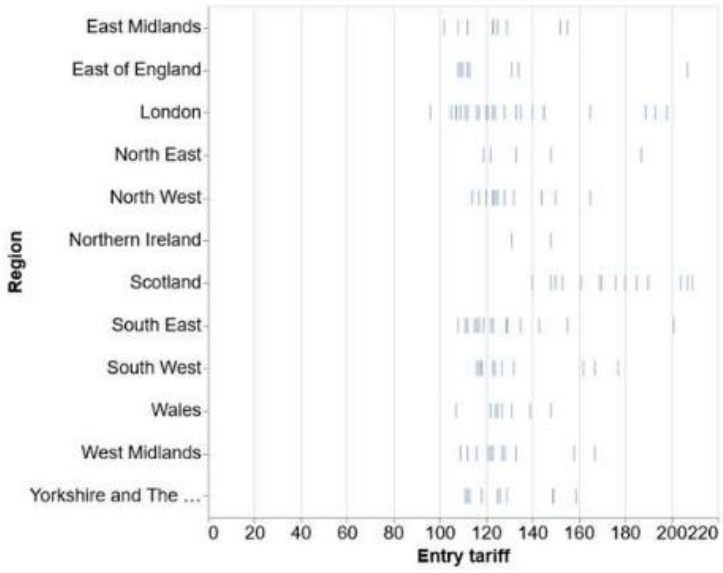
"All institution subject entry tariffs binned in histogram format"

"Create a bar chart that shows entry tariffs for UK institutions, increasing in bands of 20, and how many institutions fit within each band"

"Show me a histogram counting entry tariff by banding"

Q10: Strip plot

Institution	Region	Entrytariff	...
Anglia Ruskin	East of England	109	...
Bath Spa	South West	124	...
Bournemouth	South West	116	...
Brighton	South East	117	...
Brunel	London	123	...
...	...	...	...



"Show the entry tariff across all regions of the uk"

"Graph that shows the varying Entry Tariffs split by Region of the UK."

"Plot a line bar graph showing region vs entry tariff"

"Plot a graph to show the different entry tariffs across all regions of the UK"

"Create a graph showing the entry tariff figures per UK region"

"Create a graph depicting the entry tariffs by region"

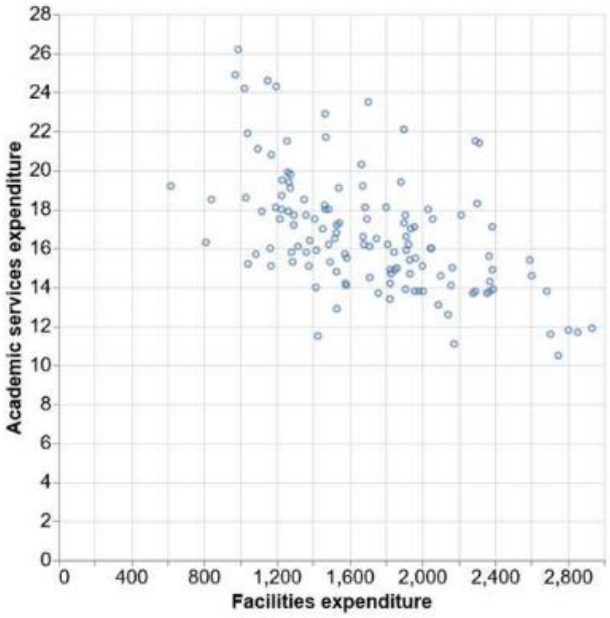
"Uk institutionsâ€™ entry tariffs by region, in some sort of line frequency graph"

"Create a chart that has different values of entry tariff on the x axis and regions across the uk on the y axis, plotting the entry tariffs for all of the institutions within those regions"

"Show me average tariff by geographic region of provider"

Q11: Scatter plot

Institution	Academic services expenditure	Facilities expenditure	...
Anglia Ruskin	19.4	1267	...
Bath Spa	16.6	1674	...
Bournemouth	19.2	1672	...
Brighton	19.1	1801	...
Brunel	17.9	1118	...
...	...	...	...



"Create scatter plot. Use expenditure on x axis and academic expenditure on the Y axis"

"Create a scatter graph to show the academic services expenditure against facilities expenditure"

"Graph that shows Facilities expenditure against Academic services expenditure"

"Show academic services expectations vs facilities expenditure"

"Plot a scatter graph to show the academic services expenditure and the cavities expenditure"

"Create a scatter graph showing the facilities expenditure over academic services expenditure"

"Create a dot graph to demonstrate the facilitate expenditure and the academic services expenditure by institution"

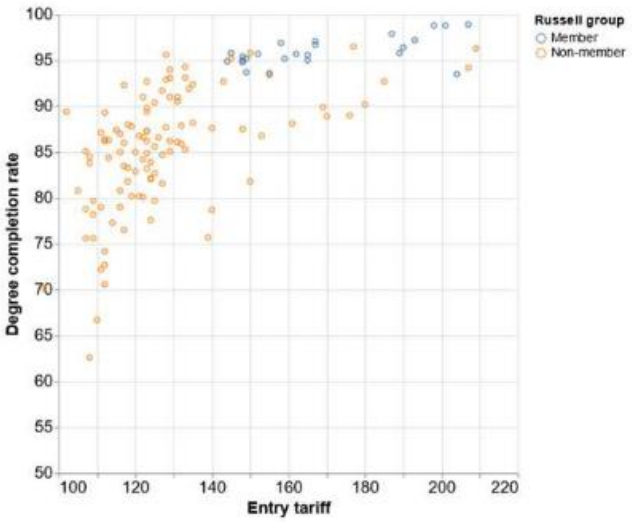
"Scatter plot for uk institutions showing academic service expenditure vs facilities expenditure"

"Create a graph which plots the academic expenditure of UK institutions against their facilities expenditure"

"Cross tab academic services expenditure against facilities expenditure using a scatter plot"

Q12: Scatter plot (coloured)

Institution	Russell group	Degree completion rate	Entry tariff	...
Anglia Ruskin	Non-member	78.2	109	...
Bath Spa	Non-member	82.1	124	...
Bournemouth	Non-member	87.0	116	...
Brighton	Non-member	86.0	117	...
Brunel	Non-member	89.4	124	...
...	...	...	...	...



"Create a scatter plot with entry tariff on the bottom and completion rate on the y axis"

"Create a scatter graph showing members and non members of the Russell groupâ€™s degree completion rate against the entry tariff"

"Graph that shows Entry Tariff and Degree completion rate, split into Russell group members/Non-members"

"Show degree completion rate vs entry tariff"

"Plot a scatter graph to show the comparison of degree completion rate and entry tariff"

"Create a scatter graph of the entry tariff against the degree completion rate"

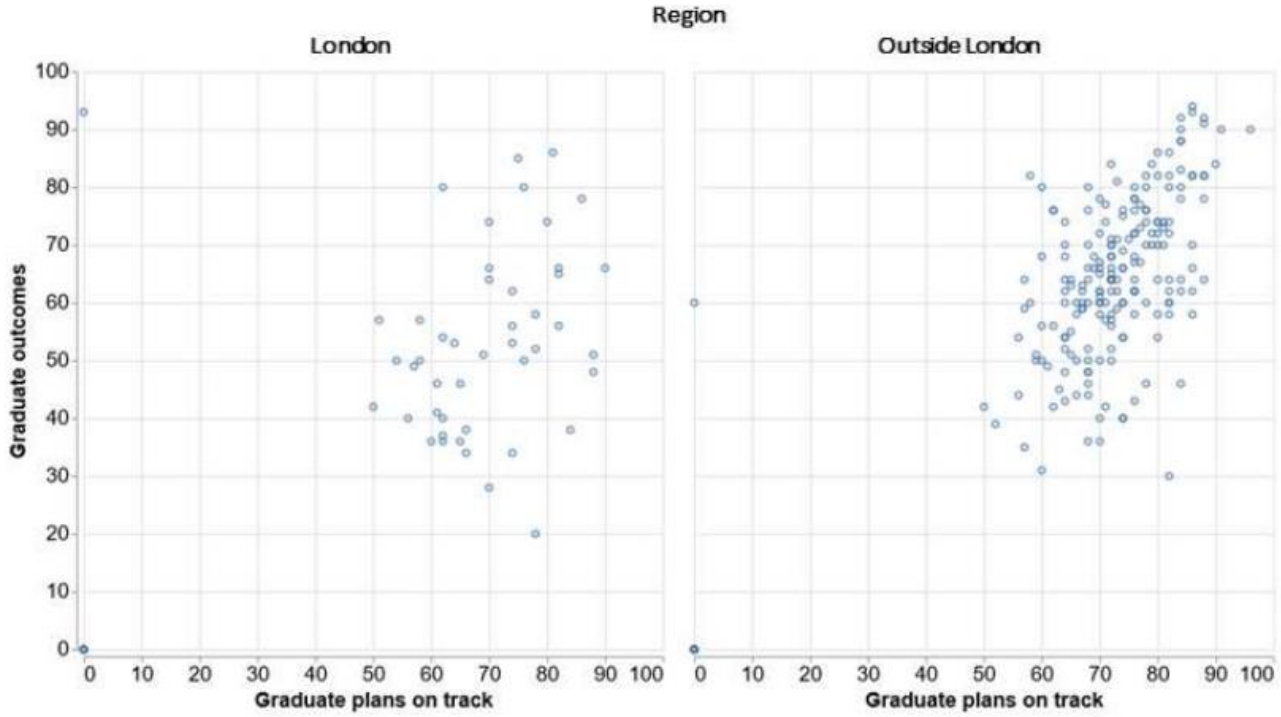
"Generate a dot graph demonstrating the degree completion rate and entry tariff for individual institutions, with different colour dots for Russell group and non-Russell group institutions"

"Scatter plot of uk institutionâ€™s entry tariffs vs. Degree completion rate further separated by Russell group membership"

"Make a scatter graph plotting the entry tariff against the degree completion rate of UK institutions, with different colours to differentiate between Russell Group members and Russell Group Non-Members"

"Cross tabulate entry tariff against degree completion rate diffentiating Russell group and non Russell Group institutions"

Q13: Scatter plot (multiples)



Subject table	Institution	Region	Graduate outcomes	Graduate plans on track	...
Accounting & finance	De Montfort	Outside London	60.0	74.0	...
Business & management	De Montfort	Outside London	56.0	60.0	...
Marketing	De Montfort	Outside London	60.0	58.0	
Accounting & finance	Loughborough	Outside London	78.0	84.0	...
Business & management	Loughborough	Outside London	88.0	84.0	...
...	...	...	...	...	...

"Create two scatter plots. One one show graduatebplans outside London on the other graduatebplans inside london"

"Plot a scatter graph to show graduate outcomes in relation to graduate plans on track both within London and outside of London"

"Graph that shows whether graduate plans are on track against graduate outcomes, separated for Regions to show those in London and Outside London."

"On individual graphs show graduate outcomes vs graduate plans on track for London and Outside London"

"Plot a scatter graph each for London and outside of London in order to compare graduate outcomes against graduate plans on track"

"Create a graph showing the graduate plans on track against graduate outcomes in and outside of London"

"Generate a graph for institutions inside and outside london demonstrating the plans and outcomes for students"

"Scatter plot of graduate outcomes vs. Graduate plans on plots separated into London and non-London regions"

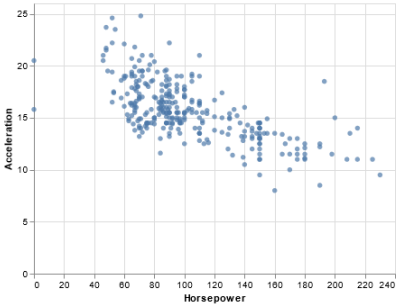
"Create two separate scatter graphs plotting graduate outcomes vs scores of whether graduate plans are on track, one for London institutions, one for those outside of London"

"Plot graduate outcomes against graduate plans on track separately for London and non London institutions"

Appendix I – Visualisations generated by GPT-Neo-125M (threshold p: 0.3, n: 3) for the in-domain cdCorpus test set

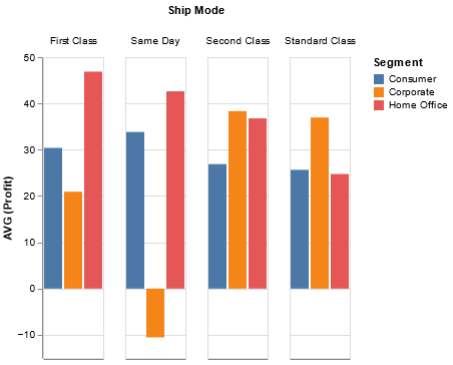
Scatterplot of horsepower vs acceleration

Visualisation:



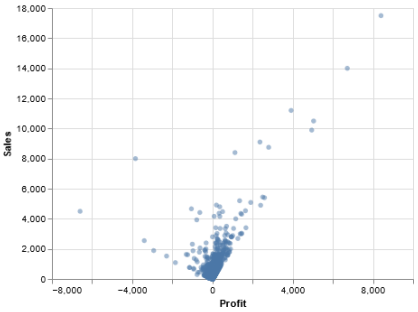
Plot IMDB rating against Rotten Tomatoes rating.

Visualisation:



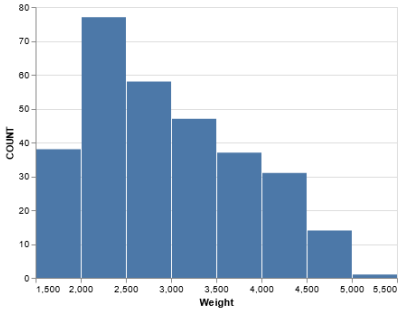
plot profit amount by sales amount

Visualisation:



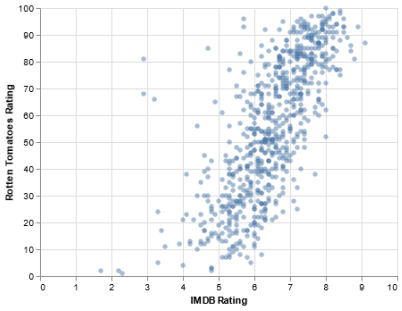
Visualize the distribution of models by weight into 8 buckets

Visualisation:



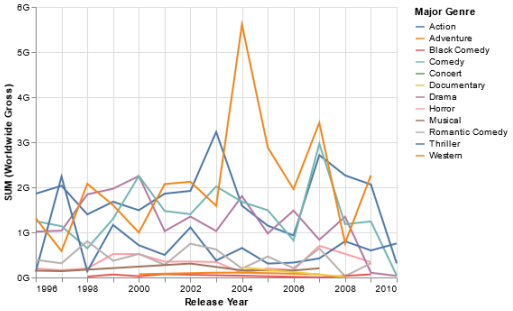
Create scatter graph of IMDB rating by Rotten Tomatoes Rating

Visualisation:



gross per genre over time

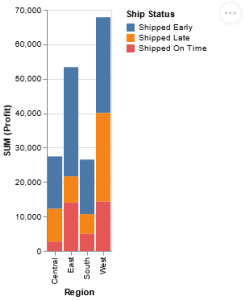
Visualisation:



Query:

Stacked columns of Sum (profit) segmented by Ship Status vs Region

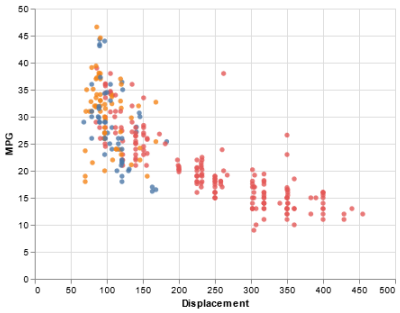
Visualisation:



Query:

Scatter plot, x axis Displacement, Y axis MPG, Color by Origin

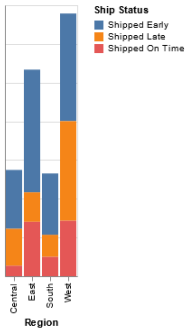
Visualisation:



Query:

What is the total profit for each region, based on ship status?

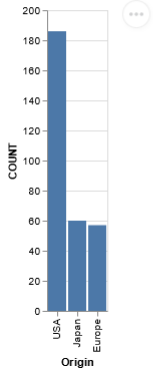
Visualisation:



Appendix J – Visualisations generated by GPT-J-6B (threshold p: 0.3, n: 3) for the in-domain cdCorpus test set

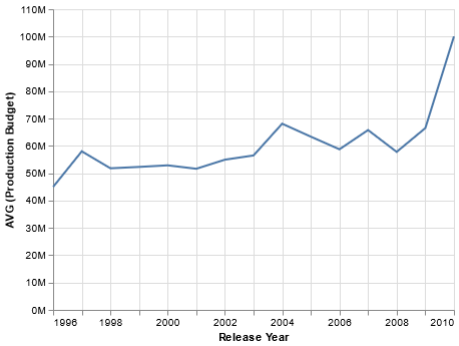
show number of cars produced by each country

Visualisation:



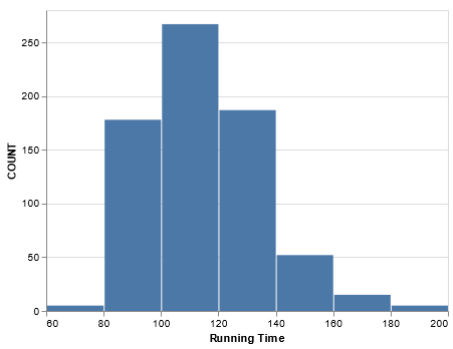
Line graph of average Production Budget across Release Year

Visualisation:



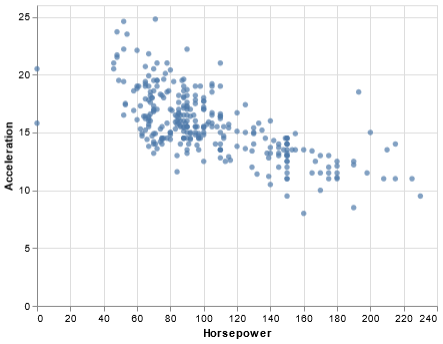
How many movies are of each length?

Visualisation:



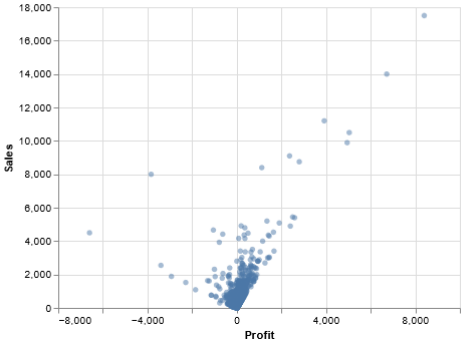
Is there a correlation between acceleration and horsepower?

Visualisation:



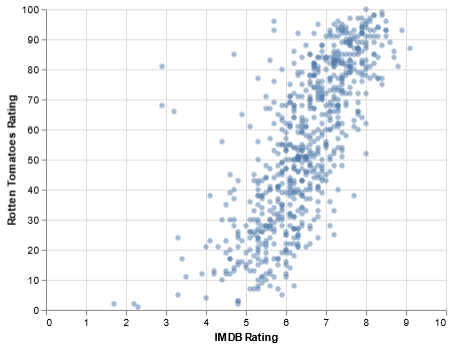
make a scatterplot of profit vs sales, with profit on x-axis

Visualisation:



Scatterplot of Rotten Tomatoes Rating by IMDB Rating

Visualisation:



histogram of Running Time

Visualisation:

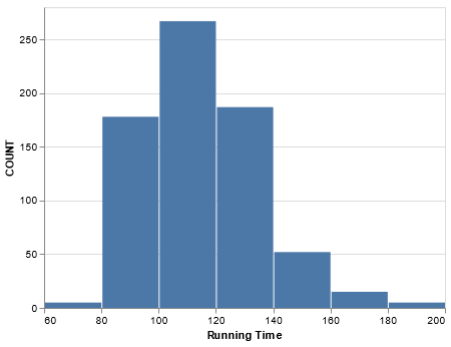
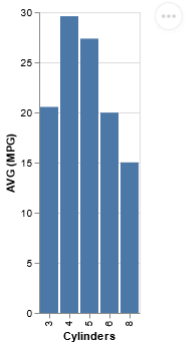


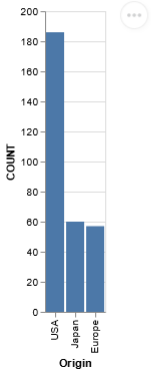
chart average MPG to cylinders

Visualisation:



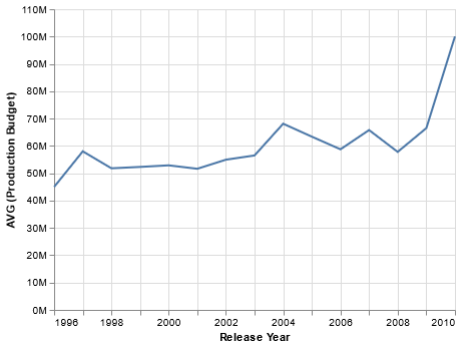
show me the count of cars by country of origin in bar chart

Visualisation:



plot average production budget over release year

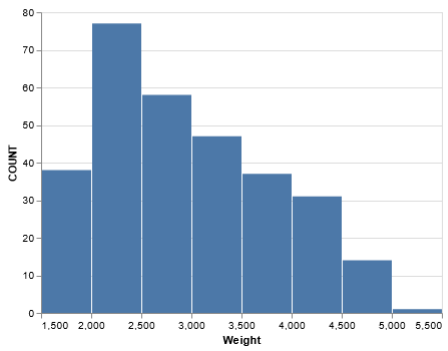
Visualisation:



Appendix K – Visualisations generated by GPT-J-6B (threshold p: 0.3, n: 3) for the out-of-domain cdCorpus test set

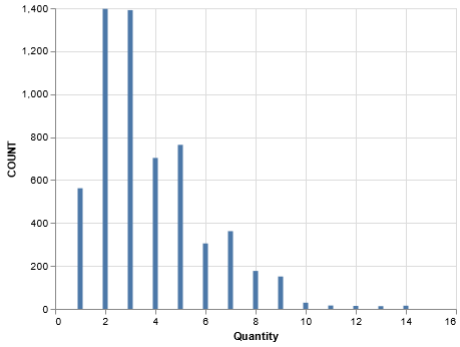
Visualize the distribution of models by weight into 8 buckets

Visualisation:



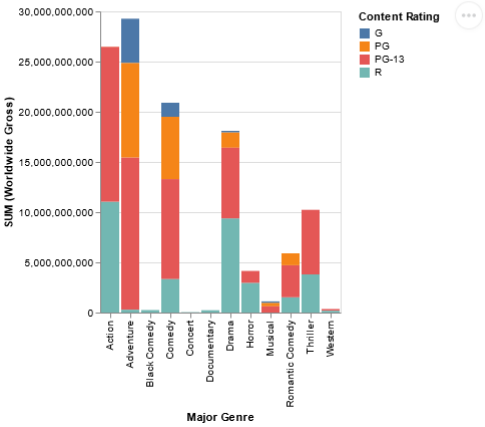
How many orders were placed for each order quantity?

Visualisation:



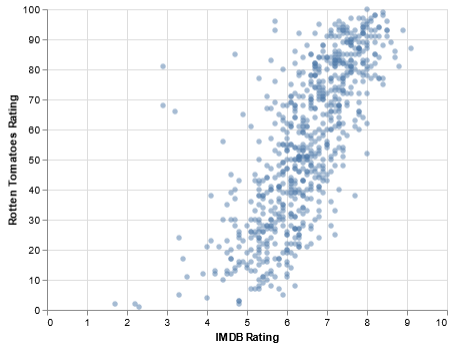
Give me a stacked column chart of worldwide gross by genre and rating

Visualisation:



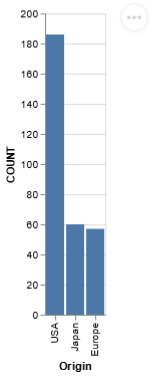
Plot IMDB rating against Rotten Tomatoes rating.

Visualisation:



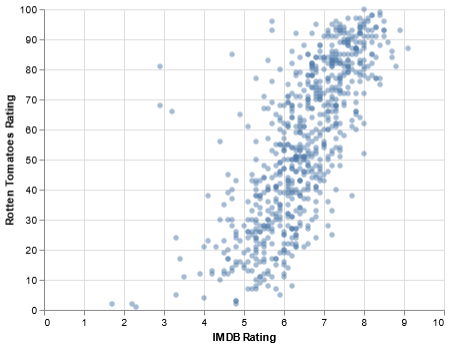
show me the count of cars by country of origin in bar chart

Visualisation:



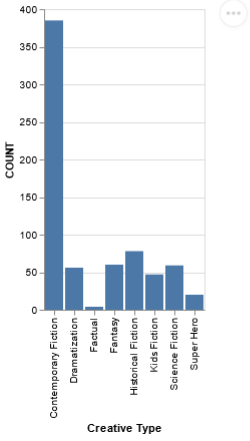
Scatterplot of Rotten Tomatoes Rating by IMDB Rating

Visualisation:



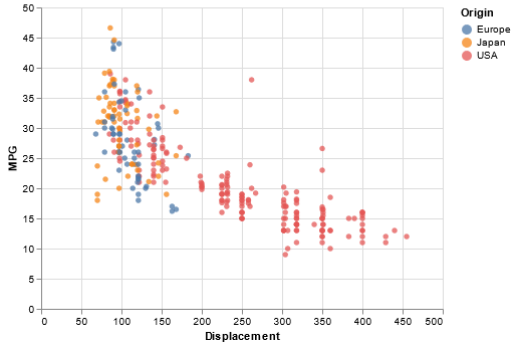
bar chart of number of movies in each creative category

Visualisation:



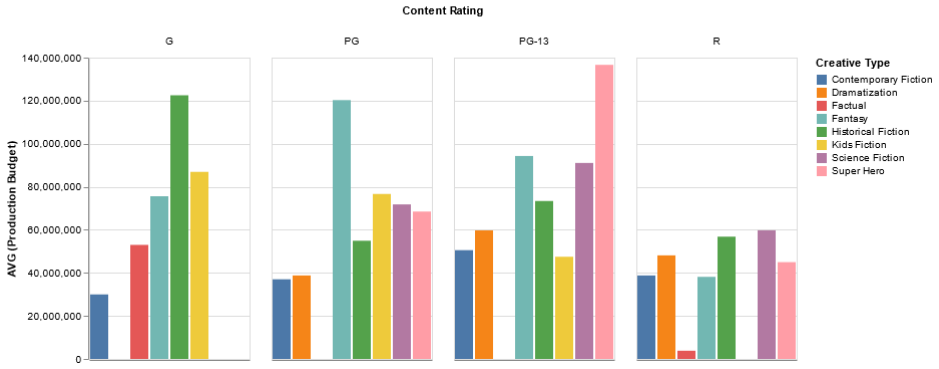
Scatter plot, x axis Displacement, Y axis MPG, Color by Origin

Visualisation:



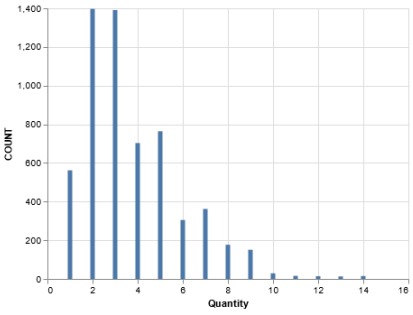
Avg(Poduction Budget) Grouped by Content Rating and sub-grouped by Creative Type

Visualisation:



Show me the count of each order quantity

Visualisation:

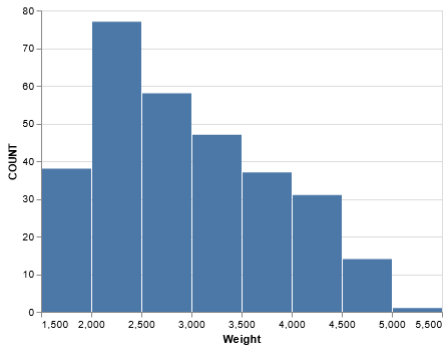


Appendix L - Visualisations generated by GPT-J-6B (threshold p: 0.3, n: 3) for the in-domain Spanish cdCorpus test set

Note: Queries are shown in English.

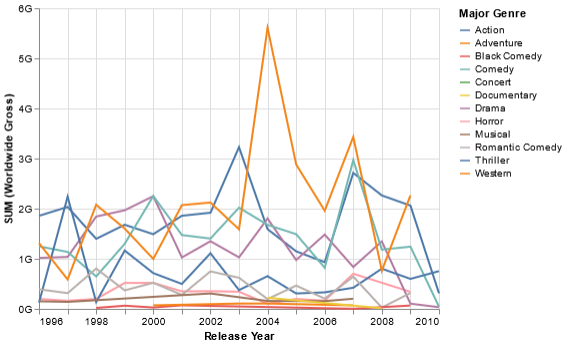
Distribution of car weight

Visualisation:



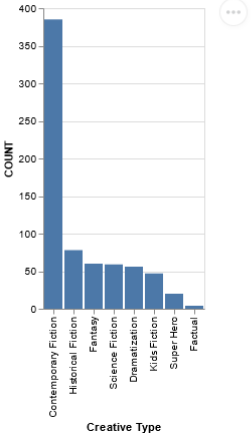
Plot the total worldwide gross over time for each major genre.

Visualisation:



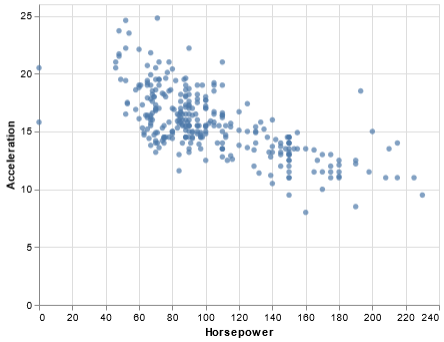
Sort creative types by number of movies

Visualisation:



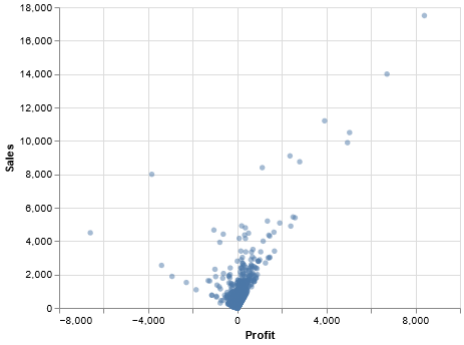
show scatter plot of accelration and horse power

Visualisation:



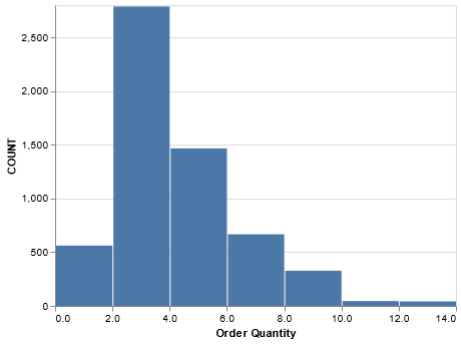
Show me a scatter chart of sales by profit

Visualisation:



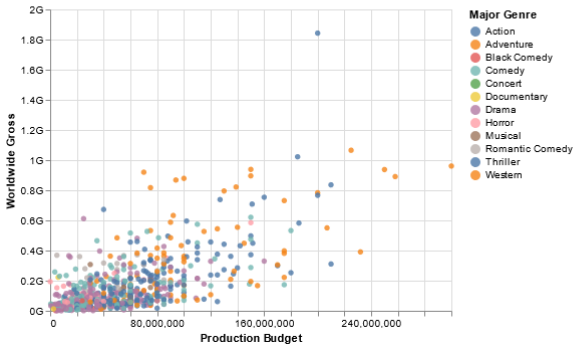
show me a bar chart of count by order quantity

Visualisation:



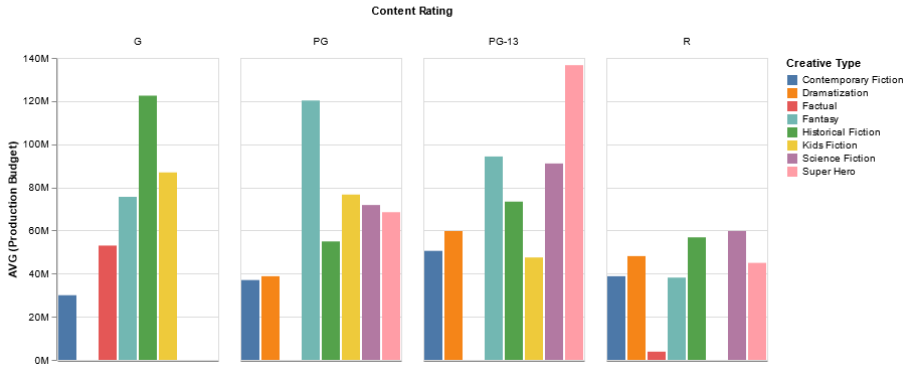
For each movie, show its budget and gross on a scatter plot and include the genre

Visualisation:



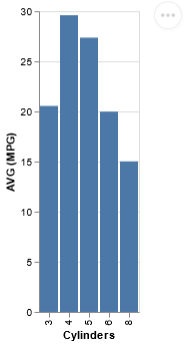
For each content rating, show a column chart of average production budget by creative type

Visualisation:



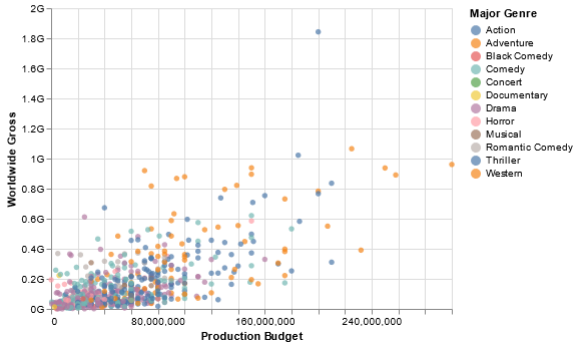
Compare AVG (MPG) with Cylinders

Visualisation:



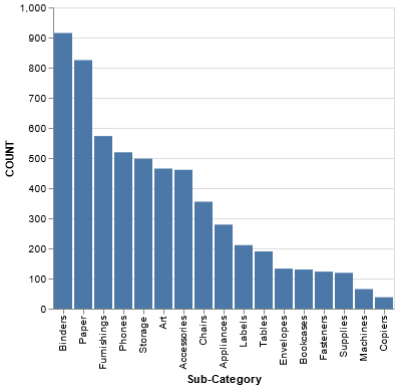
scatter plot of production budget vs worldwide gross, colored by genre

Visualisation:



Bar Chart Count by Sub-Category

Visualisation:



Appendix M - Visualisations generated by NL4DV for the Spanish cdCorpus test set

1. show me a bar chart of count by order quantity

   Spanish: muéstrame un gráfico de barras de conteo por cantidad de pedido

   Outcome: Attribute error

2. Split region by ship status and Sum(Profit) render Stacked barchart

   Spanish: Dividir región por estado de envío y representación de suma (beneficio) Gráfico de barras apilado

   Outcome: Attribute error

3. Avg(Poduction Budget) Grouped by Content Rating and sub-grouped by Creative Type

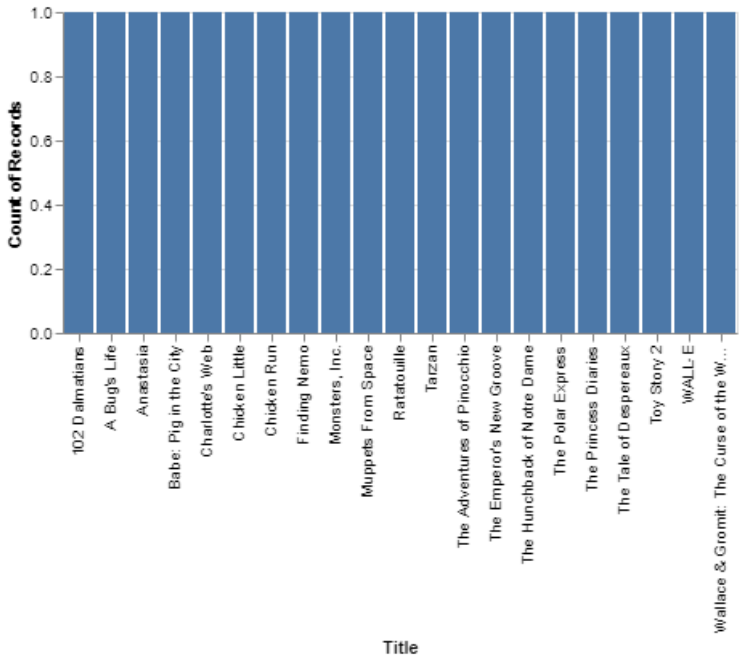
   Spanish: Promedio (presupuesto de producción) agrupado por calificación de contenido y subagrupado por tipo de creatividad

   Outcome: Attribute error

4. gross per genre over time

   Spanish: bruto por género a lo largo del tiempo

   Outcome:



5. show me distribution of gross as a strip plot
   Spanish: muéstrame la distribución del bruto como un diagrama de franjas
   Outcome: Attribute error

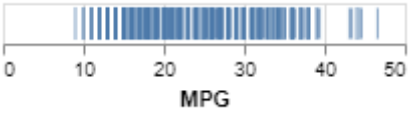
6. count by cylinders and origin
   Spanish: cuenta por cilindros y origen
   Outcome: Attribute error

7. Draw the axes for AVG(Weight) vs Year
   Spanish: Dibuje los ejes para AVG (Peso) vs Año
   Outcome: Attribute error

8. For each ship mode, plot a bar graph where each bar indicates the average profit of each segment
   Spanish: Para cada modo de envío
   Outcome: Attribute error

9. plot average production budget over release year
   Spanish: trazar el presupuesto de producción promedio durante el año de lanzamiento
   Outcome: Attribute error

10. Scatterplot of displacement vs mpg. Color by origin.
    Spanish: Diagrama de dispersión de desplazamiento vs mpg. Color por origen
    Outcome:



References:

1. Alammar, J. 2018, Jun 27,-last update, The Illustrated Transformer. Available: https://jalammar.github.io/illustrated-transformer/ [2022, December 18,].
2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. & Polosukhin, I. 2017, "Attention is All You Need", Proceedings of the 31st International Conference on Neural Information Processing SystemsCurran Associates Inc, Red Hook, NY, USA.
3. Srinivasan, A., Nyapathy, N., Lee, B., Drucker, S.M., Stasko, J. & ASSOC COMP MACHINERY 2021, "Collecting and Characterizing Natural Language Utterances for Specifying Data