C1 and C2 Project Tasks additional guidance and examples

Task guidance

The brief for this week's challenges is deliberately high level and open. Delegates are working with data sets that they are already familiar with as they had tasks in week 3 to explore, query, design and transform those data sets. The powerbi workbooks from week 3 were saved by delegates to github; it makes sense to reuse a copy of these workbooks as the starting point for this week.

The aim of this week is to build upon what delegates know about the data to try out their new python/ pandas skills AND to turn the data into analytical reports using powerbi. There are no step-by-step instructions for the tasks so we would expect a lot of variety in the output delegates produce.

Recommended sequence of steps to complete task:

- Select one of the 3 questions to work on – it doesn't matter which is selected, all the topics provide opportunities for interesting visual exploration and reporting
- **Remember to regularly back up your work to github!**
- Launch jupyter notebook, import pandas and work through your exploratory analysis of the data. Remember to annotate the notebook to explain what you are doing and what you have learnt. Notebooks can be annotated using markdown cells.
- Load data files into the notebook using pandas read, as separate dataframes
- Use pandas functionality to preview the data frames. Use the pandas user guide to help you write the commands correctly. Useful functions:
    - HEAD() TAIL()
    - INFO()
    - DESCRIBE2()
    - SHAPE
    - SAMPLE()
    - ISNULL().SUM()
    - NUNIQUE UNIQUE()
    - COLUMNS()
    - DUPLICATED()
    - DTYPES
    - CORR()
- Consider using pandas functions for joining the data frames, filtering the data frames, reshaping and summarising the data frames. Applying these techniques may help you explore the data visually in the next step.
- Use pandas visualisation and separate charting libraries like plotly, seaborn and matplotlib to explore the data visually.
    - Typically your EDA workflow will involve examining one column at a time to see the distribution, gaps and outliers before plotting 2 or 3 columns together using appropriate chart types like catplots and scatterplots to

understand relationships, compare trends, before, finally, creating multi plot visuals like pairplots/pairgrids, correlation matrix heatmaps to review all the columns in one go. For this you might find the individual visualisation library documentation helpful or you can look at EDA notebook examples in Kaggle for inspiration. With a little internet searching you may discover there's a couple of great EDA libraries available in python that could be installed and deployed.

- Once you have explored the data in python/pandas you should have a better idea of what you will show in your report. Open the existing powerbi workbook and review the relationships and transformation steps already created
- Brainstorm the ingredients of the question – what data points are needed? what transformation / cleaning steps might be needed? Will you need new columns?
- Create a rough sketch (using pen & paper) of the report you can imagine building. You can use the below questions to help you think this through:
    - What chart types will be most useful for this report? Think about what kind of data insight you are trying to show or what message you are digging into. Keep in mind that powerbi has a limited number of visual types out of the box but you can add more visuals to your workspace if need be.
    - Whats the most important information that you will seek to highlight in the report? That's probably information you want to put in your title, or on the very top of the report view.
    - Does it make sense to include filters / slicers to enable focus in the report onto certain categories or to enable comparisons?
    - What design decisions would you prioritise to make the reporting clear and accessible to someone else? – e.g., colour choices, themes, interactions between reports, instructions, title and chart headers, use of whitespace, tooltips
- Start compiling the powerbi report for your chosen question, returning whenever needed to the transformation/ power query editor window to transform the data.
- Consider if you need multiple report pages to share the information you have found – e.g. a top page summary and more detailed reports underneath.
- Once you have a report ready – give it to someone else to test.
- Now, get ready to share and show your work to the whole class on Friday!

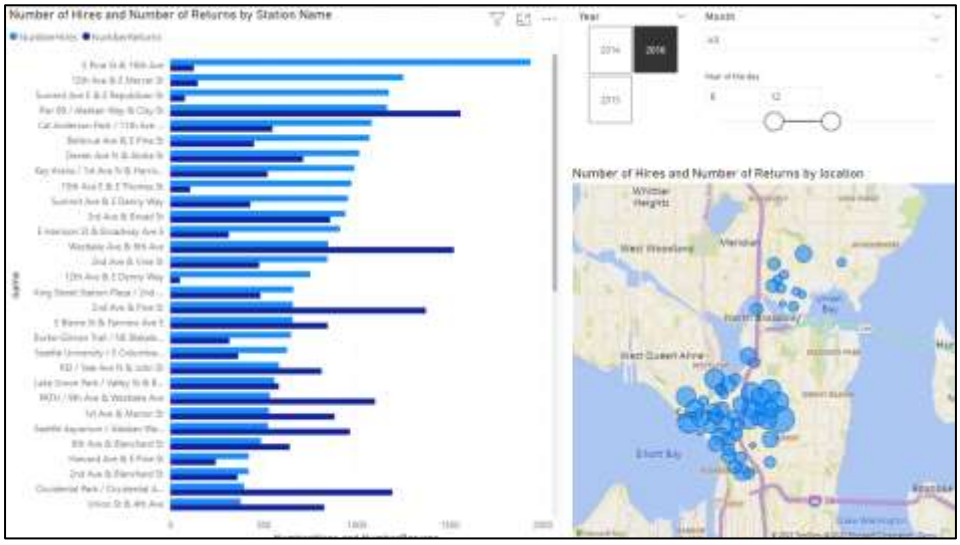
Example reports (no need to reproduce; these are just for inspiration!)

C1 – 1

Which are the most popular Seattle cycle hire stations (either for hiring a bike, returning a bike, or both)? Does the popularity of some cycle hire stations vary by month or hour of the day? Are there some cycle hire stations which have a lot of bikes hired from them but very few dropped off at them (forcing organisers of the scheme to manually transfer bikes between stations) ?

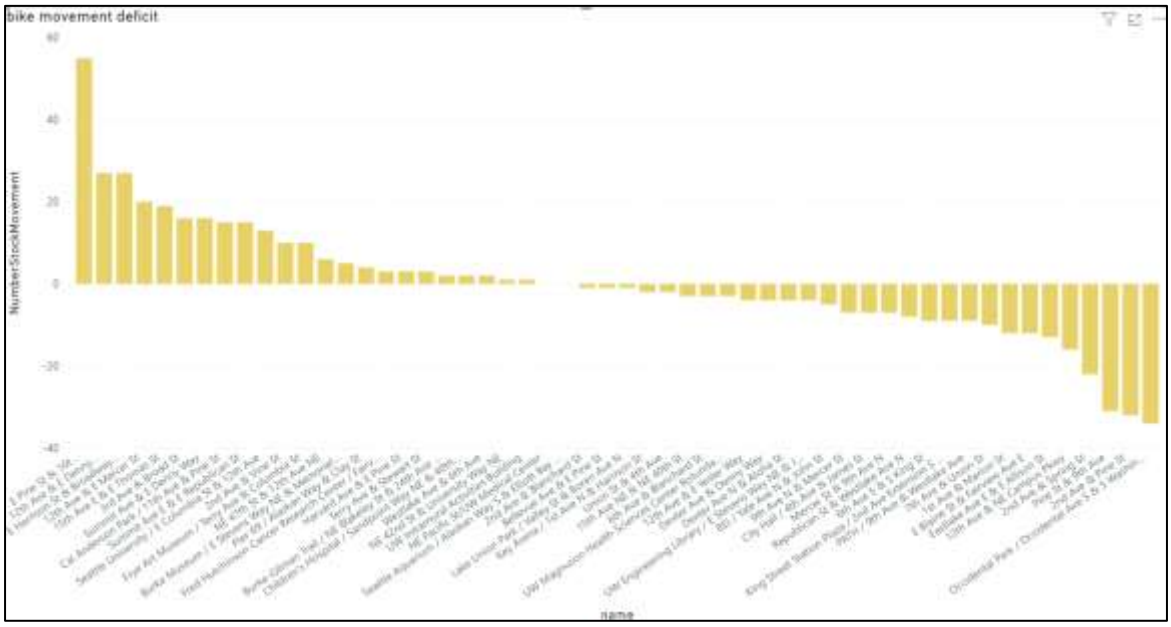
This hire and return report was created by:

- Connecting the trips to the stations with the to_station_id (optional relationship)
- Calculating a measure for the number of hires total
- Calculating a measure for the number of returns to each station using the relationship



This bike movement deficit chart was created by:

- calculating the difference between the number of hires from and no of returns to a particular station, plotting them on a chart

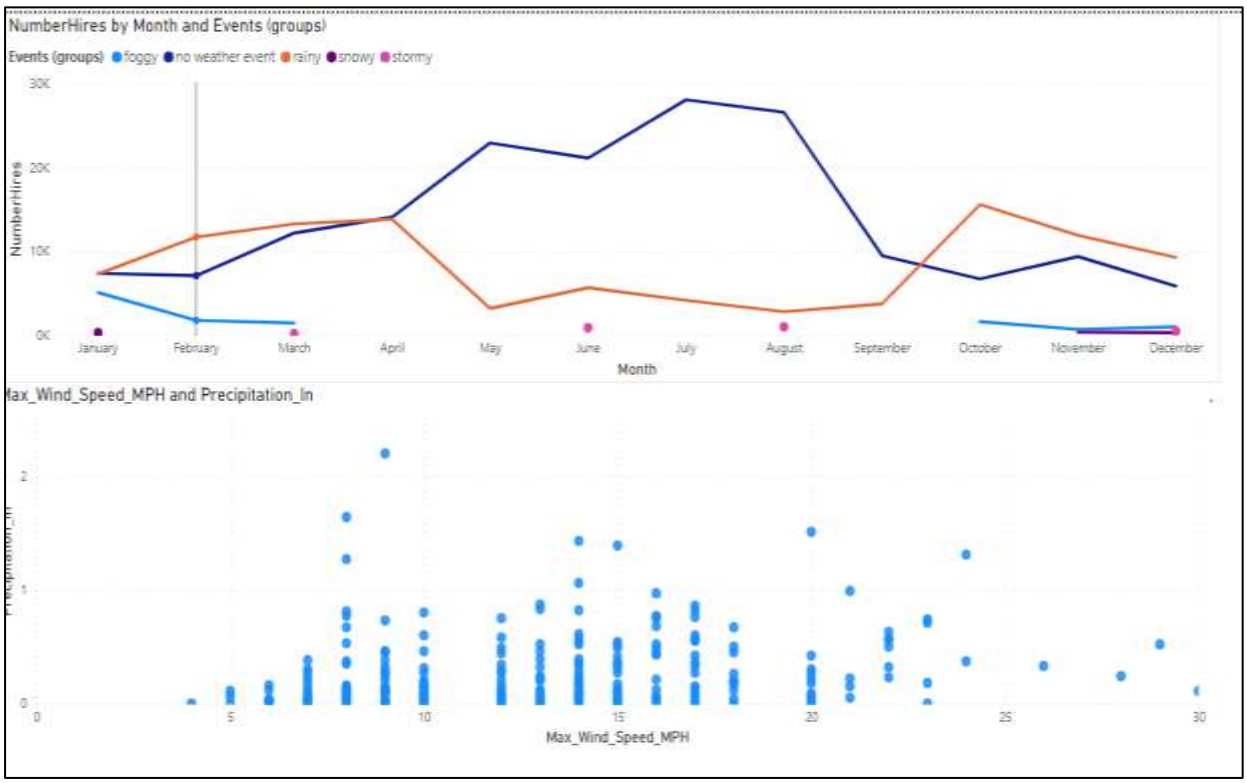


C1 – 2

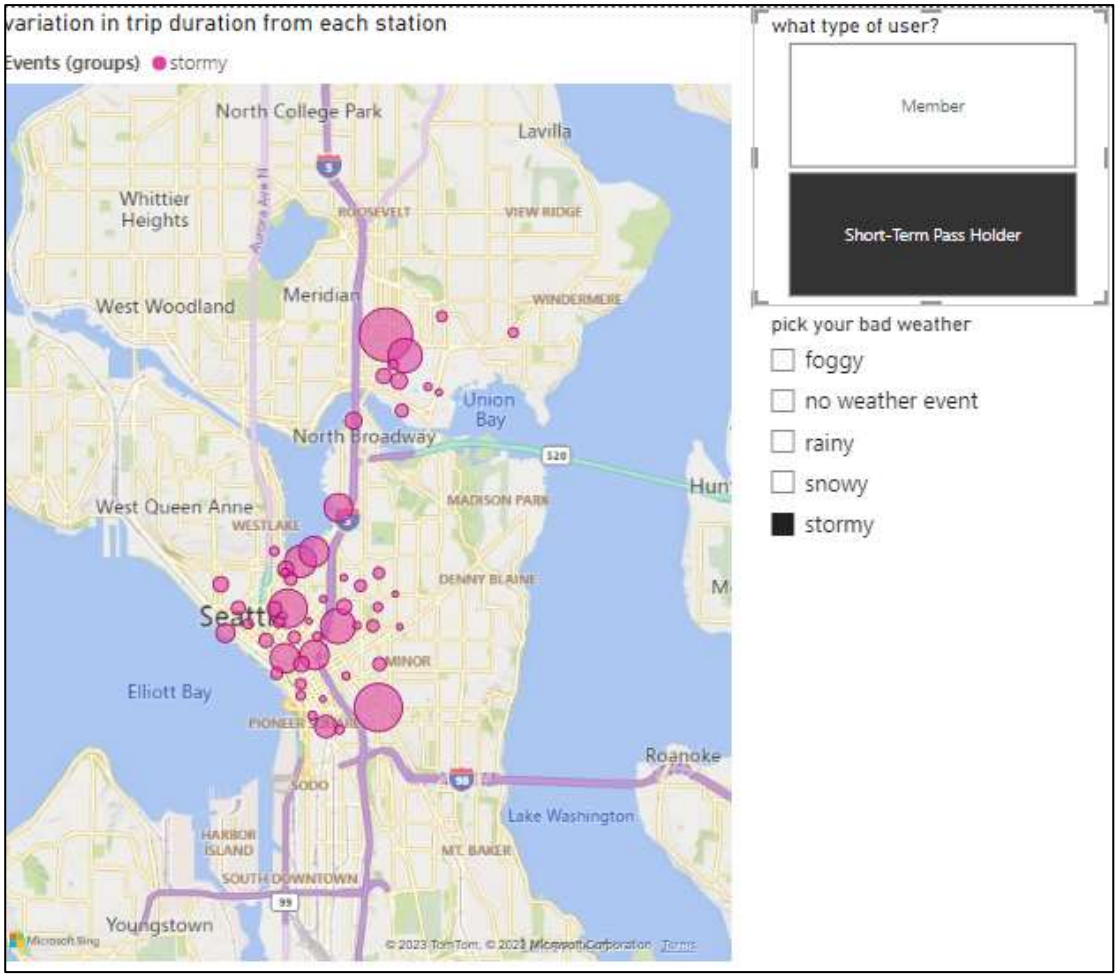
Do Seattle cyclists hire bikes at the same average frequency and ride approximately the same average distance in all types of weather conditions or is there evidence to suggest that some weather conditions are more cycle hire friendly than others? Does this trend vary during the year or by neighbourhood?

This exploratory weather versus trips report was created by

- grouping the weather events
- showing the number of hires on a time series
- comparing weather information on a scatter which can be filtered by the line chart



This trip duration variation report was created by using the same weather category as above, and filtering out the normal days

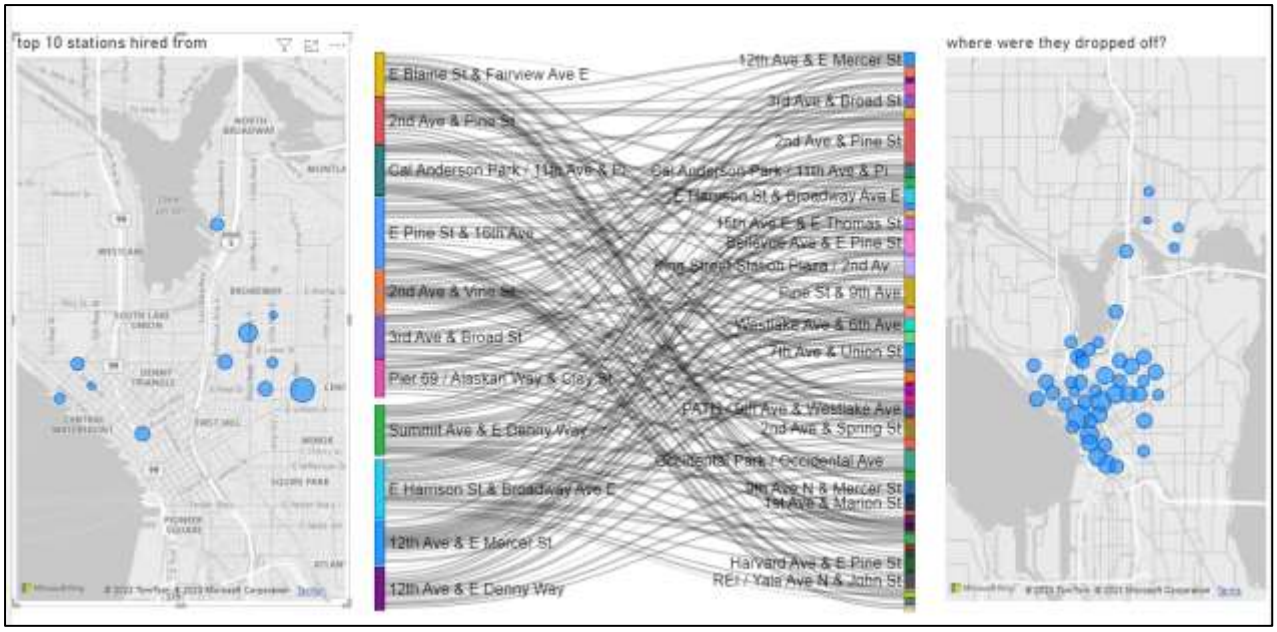


C1 – 3

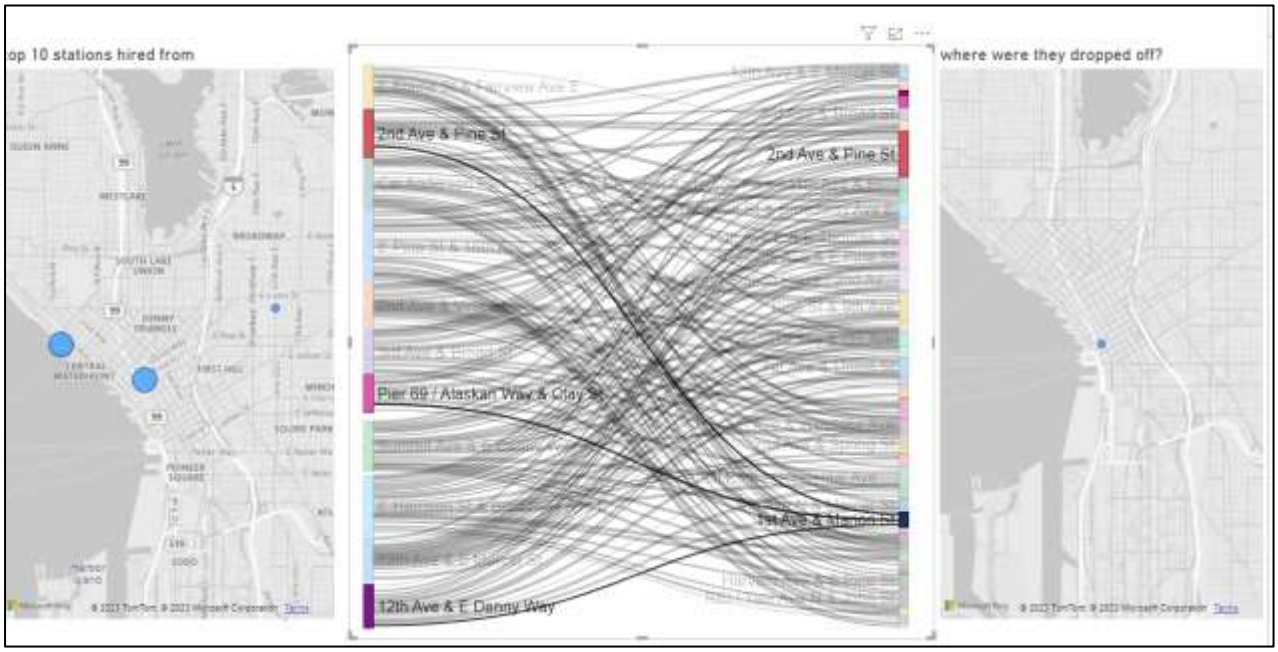
Are there observable patterns in the neighbourhoods people tend to hire cycles from and where they cycle to? Can we see a commuter trend, for example, or evidence of students hiring cycles in term time? Are there any outliers from the 'usual' journey that you spotted, and do you have any plausible explanations for this – such as where clusters of businesses are located or traffic jammed streets?

This top 10 hire from Sankey / map was created by:

- Filtering the charts by top 10 stations hired from
- adding a Sankey visual type to powerbi
- adjusting the interactivity of the charts



Example : bikes dropped off at 1st ave and marion street station



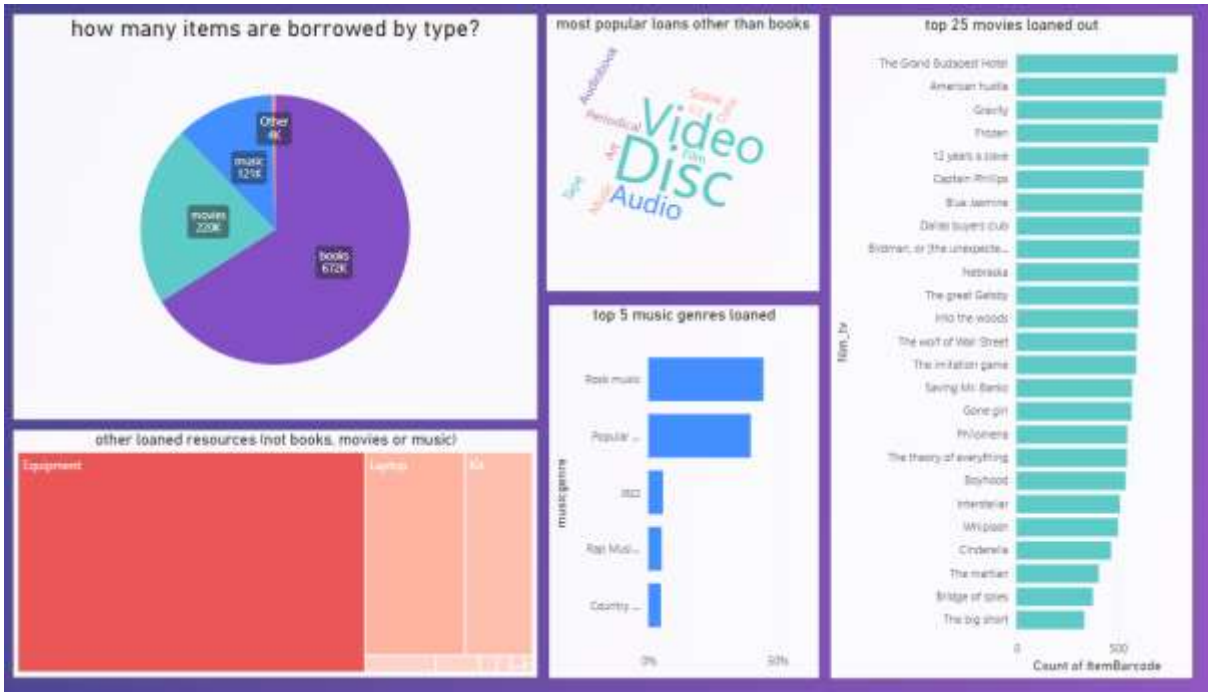
The next step would be to find out more about whats happening at or near those locations!

C2 – 1

What are the most popular types of items and collections checked out of the Seattle libraries ? Is there a particular subset within items or collections which is most frequently checked out? Are you surprised by the variety of items the library loans? Can you create an engaging visual which reveals what Seattle's libraries have to offer beyond books?

The borrowing by type overview was created by:

- Defining a high-level type using a grouping method on item description
- Using this grouping to filter the other charts
- Duplicating subject, title, and description columns, applying splits on the various delimiters to find a movie title, music genre and description of item



C2 – 2

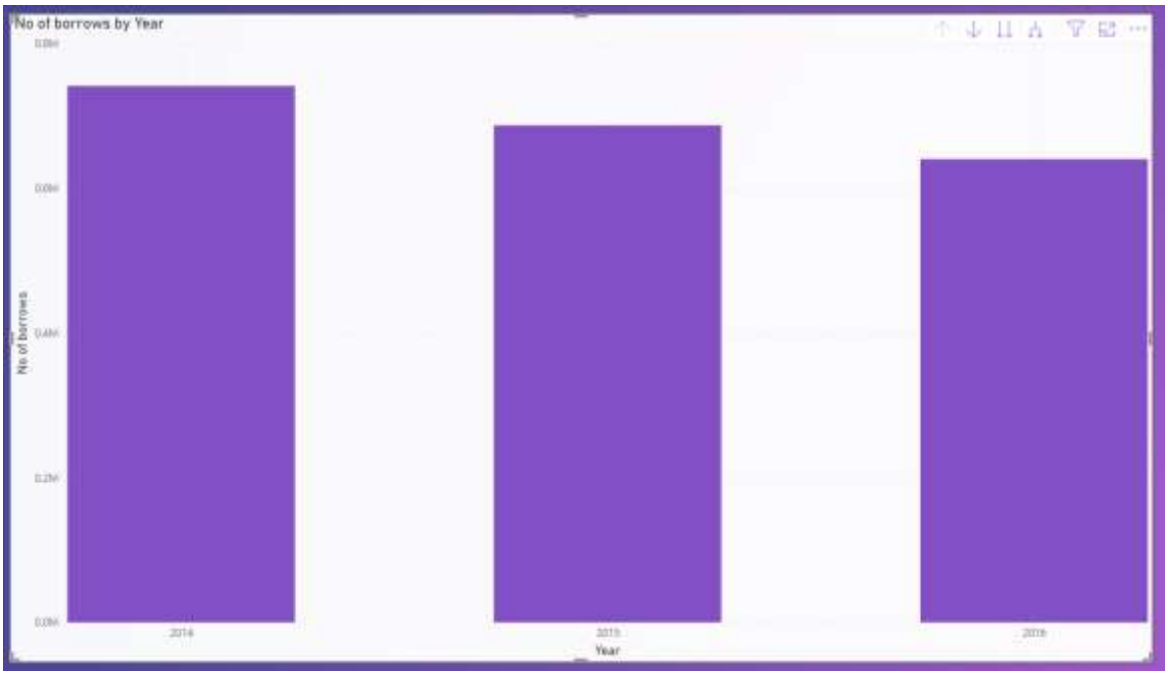
What dates or times of the year do you notice more checkouts occurring from Seattle libraries? And does this vary by media type; for example, is there a day of the week on which DVDs are typically checked out from the libraries? Can you plot a few time series visuals to show the patterns in lending, indicating when Seattle's libraries will need the most staff on hand to help borrowers?

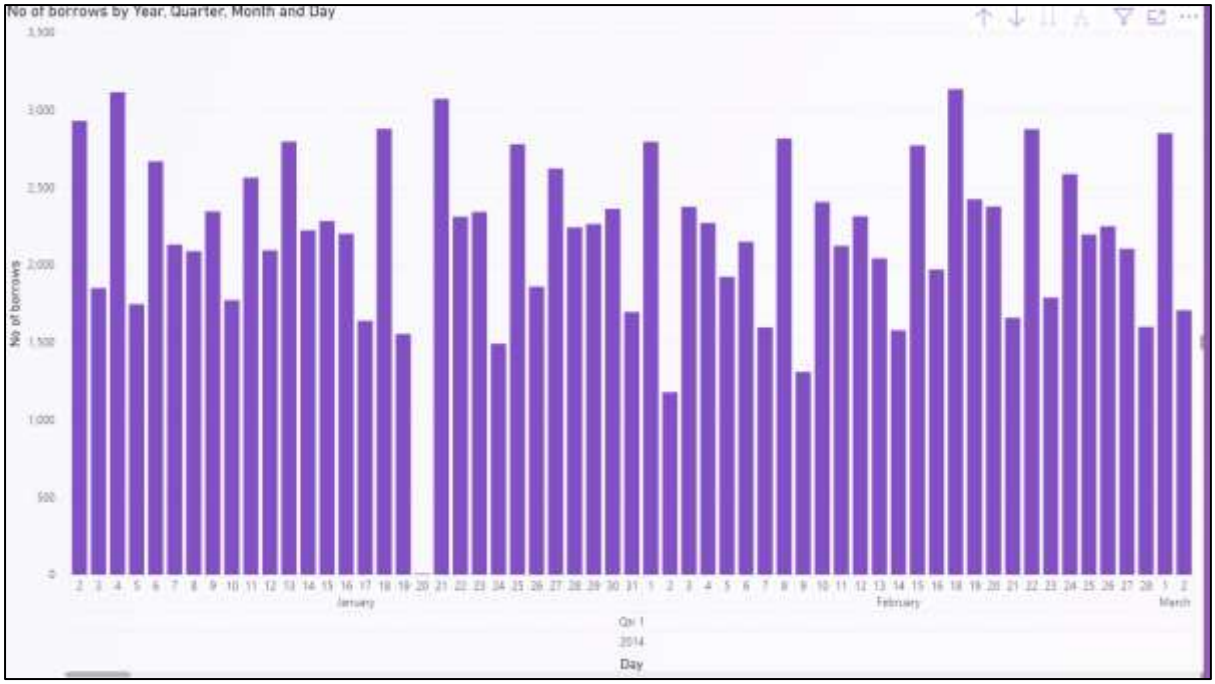
The overview of borrowing by day – hour - month report below was created by:

- Adding hour and weekday name columns in the transformation window
- Utilising slicers with two fields in each, to explore categories and formats as a drill down. The slicers have been configured to remove blank values.

The drillable date explorer below was created by:

- Making the most of the date hierarchy in the checkouts table, using the drill up and down features of the bar chart report

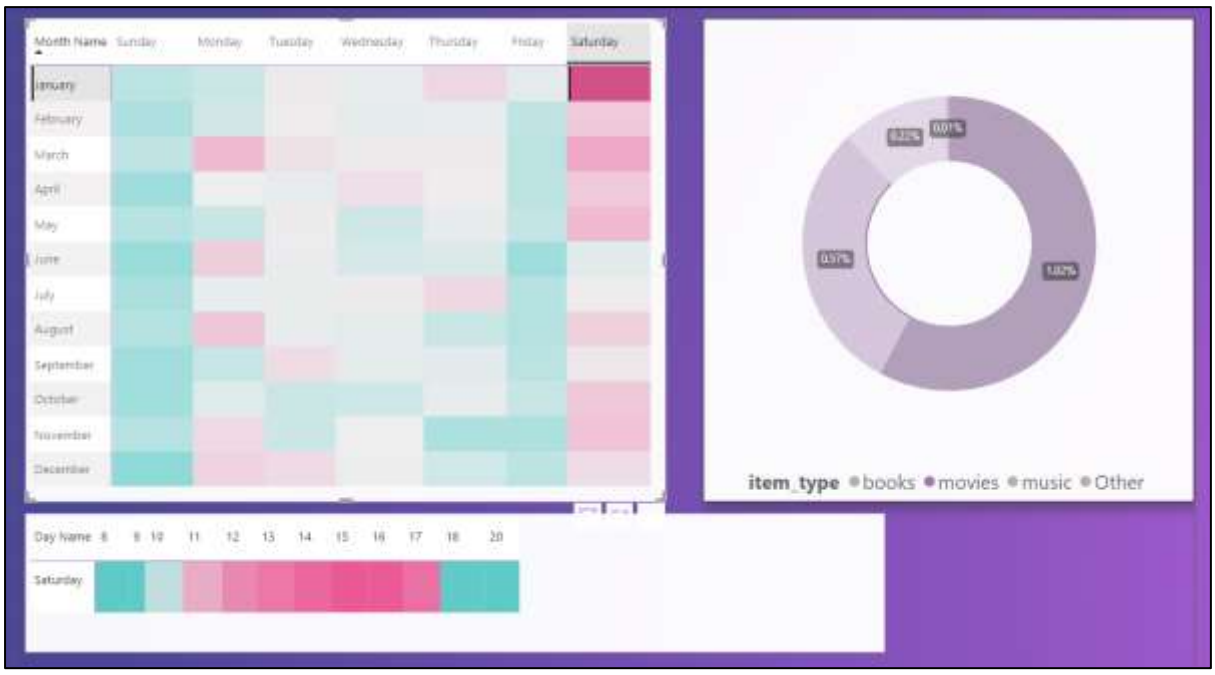




The interactive matrix visuals with media type donut chart below are created by:

- Adding month and day of week numbers as columns for sorting month and day of week names (in data view)
- Using conditional formatting in a matrix visual type and matching the background and text colours
- Creating a grouping and adding a donut with media type

Example : Saturdays in January are busy for movie borrowing. The busiest time is 3-4pm



C2 – 3

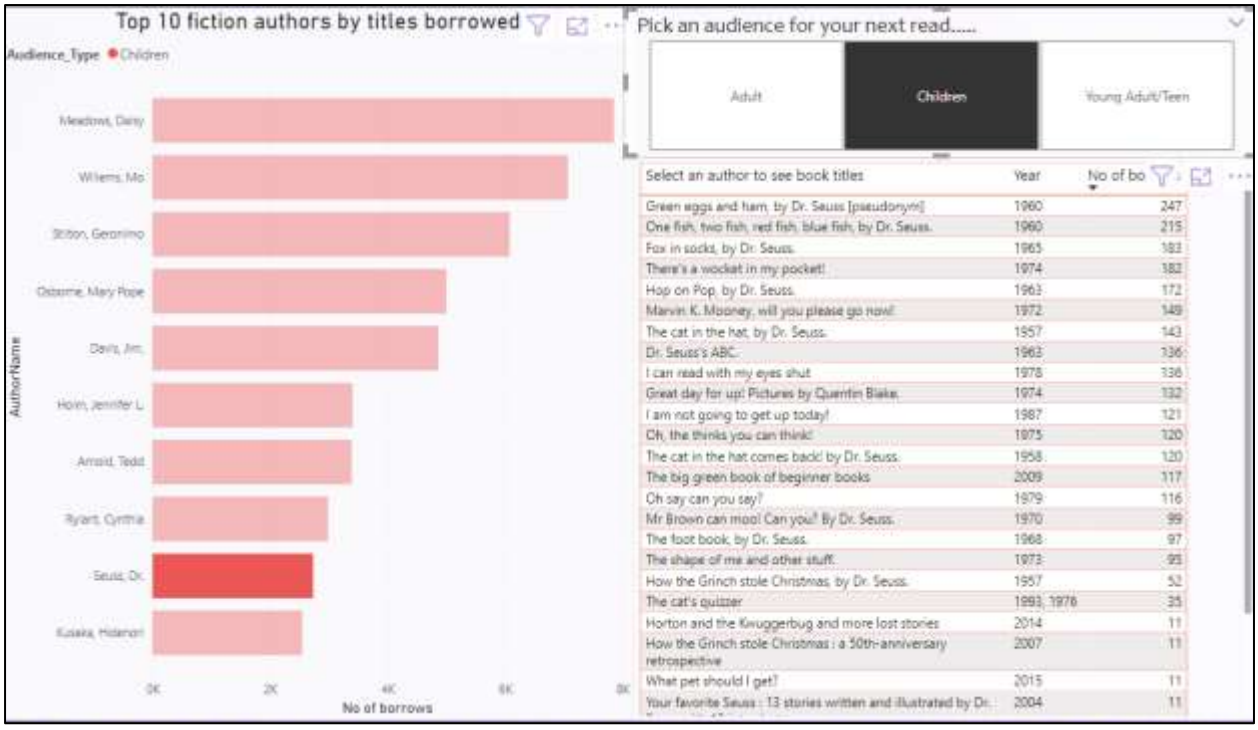
Who appears to be Seattle's favourite fiction authors in our data sample? Drill into the top 10 authors of books checked out in the collections for young people/ children or adults. Can you create an interactive visualisation to explore popular fiction authors that someone new to the library could explore to pick their next read?

The report created and shown below is achieved by:

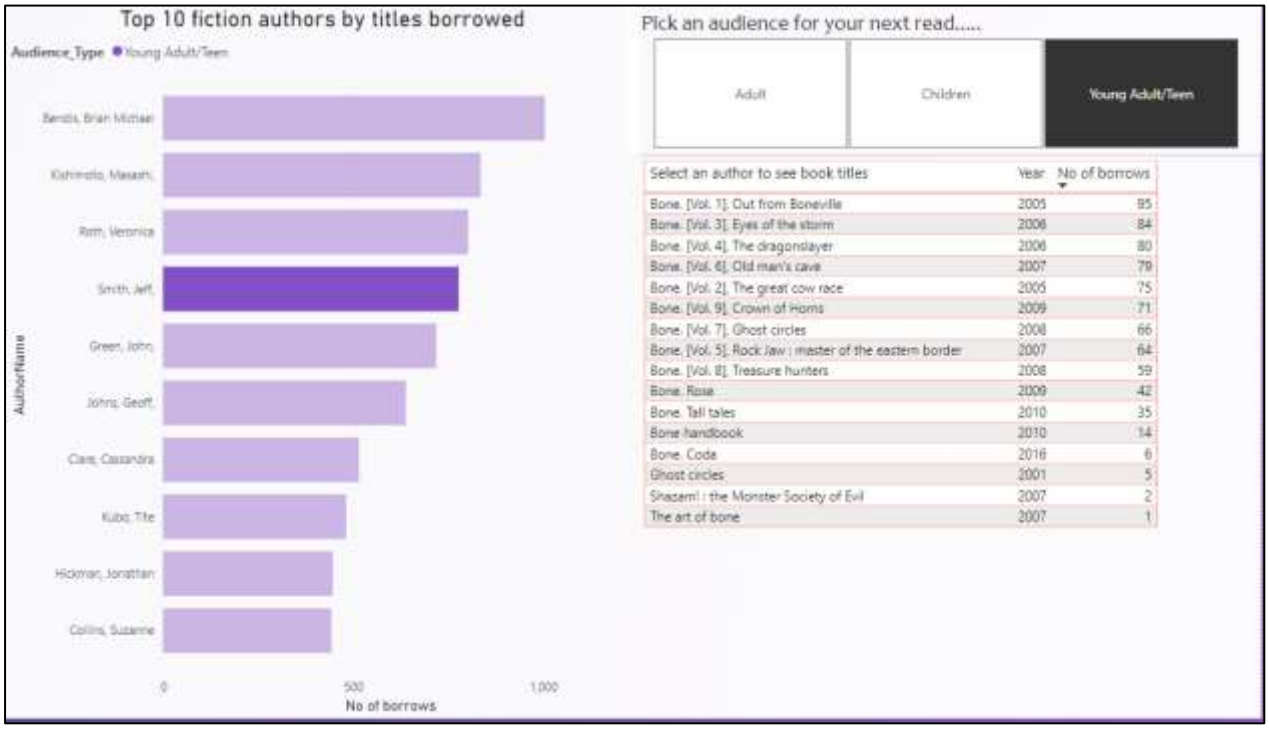
1. cleaning author, title and publicationyear in the inventory table
2. creating a conditional column based on collection.description to assign titles to Adult, Children or YoungAdult/Teen audience based on the keywords in the description
3. using the count of checkouts.itembarcode as number of borrows
4. filters from the collection table on the page – we are only interested in: Fiction/Print /Books
5. filter from the inventory on the page – author is not blank or empty

As the screenshots below show, the report allows for selection of an audience (adult, children or YA/Teen) to reveal the top ten most popular authors by borrows at Seattle's public libraries. If one of the top ten authors in that audience type is highlighted then their titles, year of publication and the number of times the book has been borrowed will be revealed. The titles are sorted by popularity (number of times that title has been checked out) and the authors are also sorted similarly.

Example 1 : children's authors > Dr Seuss



Example 2 : YA/Teen authors > Jeff Smith



Example : Adult > Louise Penny