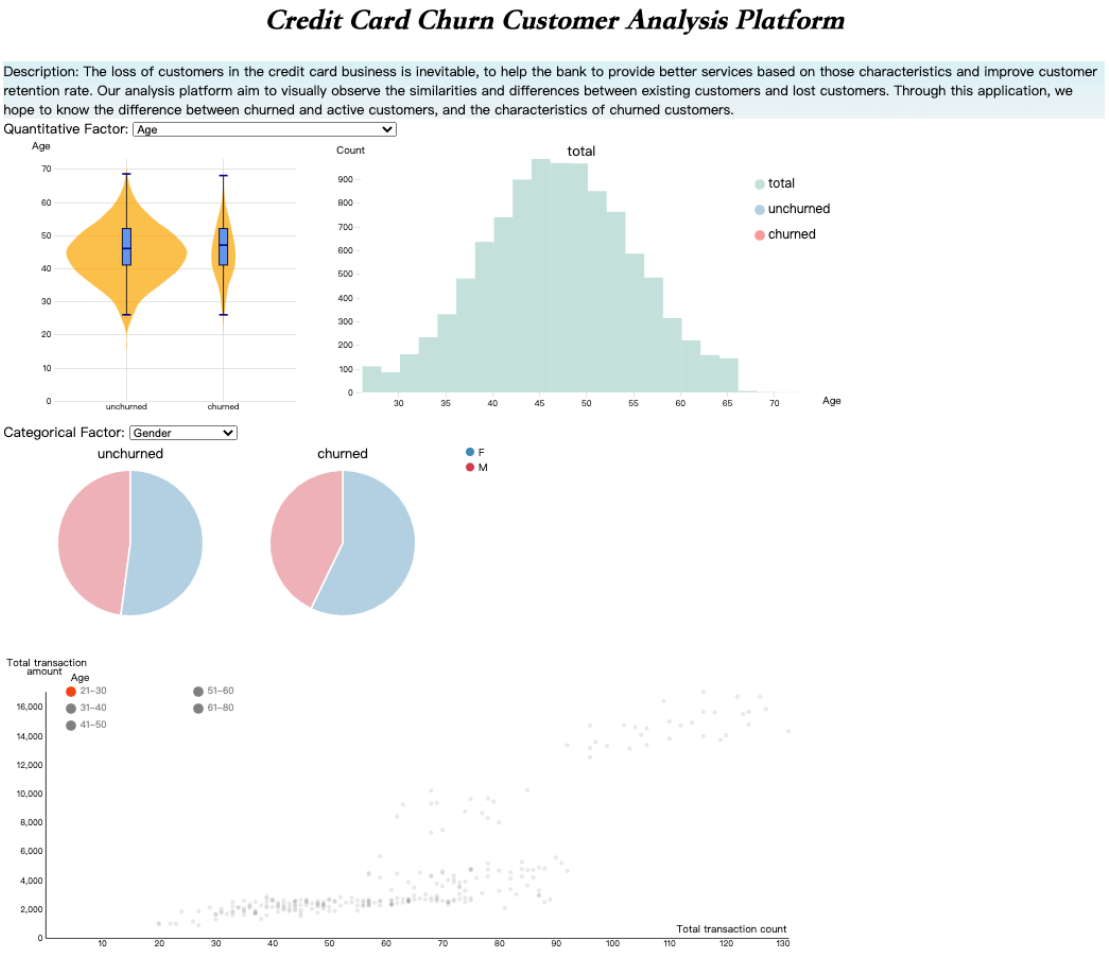
Group 23

Final Report

1. Overview

screenshot of your visualization.



Concise summary of this project

The loss of customers in the credit card business is inevitable, to help the bank to provide better services based on those characteristics and improve customer retention rate. Our analysis platform aim to visually observe the similarities and differences between existing customers and lost customers. Through this application, we hope to know the difference between churned and active customers, and the characteristics of churned customers.

To realize our goals, we have 4 views in the visualization, churned and unchurned customers are always clear to differentiate, categorical and quantitative attributes are displayed separately with their different corresponding interaction. We can tell the difference in boxplot and violin plot

between two types of customers and see the detailed distribution of the selected quantitative category in histogram; We can tell the difference in pie chart between two types of customers and see the detailed lost customers distribution and trend of the selected and highlighted categorical category with different age level in scatterplot; The type of customer selected by pie chart and boxplot is always consistent.

2. Data

The dataset we used is a table dataset. Each item represents a customer with detailed personal and credit card information. To visualize the similarities and differences between churned and unchurned customer, we extracted the causal factors of categorical type and quantitative type for each those customer.

[Note: All listed attributes are not duplicated (mutually exclusive).]

View 1 Pie chart

We derive the two groups of sample (churned and unchurned) to show their difference.

The attributes visualized in the pie chart are all **categorical** including

- Gender (2 levels, female and male),
- Education level (7 levels),
- marital status (4 levels),
- Income category (6 levels) and

We add one more derived **quantitative** attribute here

- Percentage of selected part in a whole pie(range[4, 57]) [derived].

View 2 Box plot + Violin plot

We derive the two groups of sample (churned and unchurned) to show their difference. Within each group, five quantitative attributes representing original attribute's distribution is derived for boxplot. There are one derived ordered key attribute (bin) and one derived quantitative value attribute (item count per bin) for violin chart.

The attributes visualized in the box plot and violin view are all **quantitative** including

- Age (range[26, 68]),
- Dependent Count (range[0,5]),
- Months on Book (range[13, 56]),

- Total Relationship Count (range[1, 6]),
- number of months Inactive in the last 12 months (range[-0, 6]),
- number of contacts in the last 12 months (range[0, 6]),
- Credit Limit (range[1.44k, 34.52k]),
- Total Revolving Balance (range[0, 2.52k]),
- Average Available Balance of last 12 months (range[3, 34.52k]),
- Change in transaction amount (range[0, 1.49])]),
- Total Transaction Amount (range[510, 0.58k]),]),
- Total Transaction Count (range[10, 94]),]) and
- Average card utilization ratio (range[0, 1.0]).

View 3 histogram

The attributes visualized in the histogram view are all the same as that of view 2 (box plot + violin plot). There is one derived ordered key attribute (bin) and one derived quantitative value attribute (item count per bin).

- Count of Customer (range[0, 4515]) [derived].

View 4 Scatterplot

This visualization used the derived attribute the average transaction amount. The x-axis represents the total transaction count (shown in view 2: box plot + violin plot) and the Y-axis represents the total transaction amount. The average transaction amount is showing in the tooltip with other information.

- Average transaction amount (range[19.1, 190.2]) [derived] .
- Total transaction amount ([510,10583])
- Total transaction count([10,94])

Include a URL linking to the source of the data.

Link to the source of data:
https://www.kaggle.com/sakshigoyal7/credit-card-customers

Briefly describe your current data preprocessing pipeline, if there is one.

Dataset used in the current project can be found in data folder(./data/). No data preprocessing.

3. Goals and Tasks

View 1 Pie chart

Our goal for the pie chart view is to allow users to explore and compare the behavior of churned for each categorical factors. The similarity of the two types of pie charts suggests that the chosen attribute has little effect on the loss of customers. Besides, Two types of customer can be selected and the selected type(churned or unchurned) should always be consistent with view 2.

View 2 Box plot + Violin plot

Our goal for this box plot and violin plot is to allow users to explore and compare the behavior of churned for each quantitative factors. It's innovative in an overview-detail way and linked with the histogram to show the more detailed distribution with specific axes values. So we can discover the distribution and would like to know which factors are likely to cause the difference among the customers.

View 3 histogram

In this histogram view, we aim to discover distribution and see clearly which bar has the most frequency according to the quantitative factor chosen in view 1 (pie chart). It's helpful to know which types of customers are most likely to be lost, meaning that services should target them accordingly.

View 4 Scatterplot

In this scatterplot view, we aim to show the average transaction amount within different age levels. Users can find the correlation between total transaction amount and total transaction count, and locate the outliers to take targeted operation.

4. Visualization

Our visualization can successfully realize the aimed function. Firstly, the differences between existing and lost customer are clearly demonstrated through categorical factor and quantitative factor. Then, for each type of factor, there is a more supplementary detailed view. Histogram supplements boxplot and violin plot, scatter plot supplements pie chart.

View 1 Pie chart

Encoding:

Channel angle encodes the number of items within the specific level.

Channel color hue encodes a level in the category.

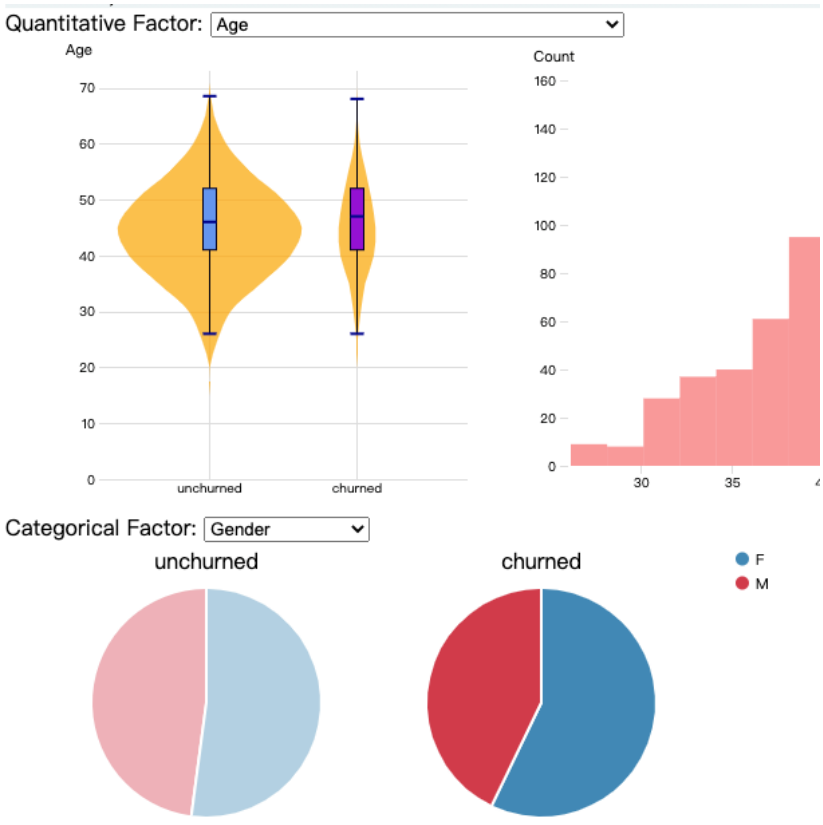
Mark of type line encodes levels in the category.

Rationale:

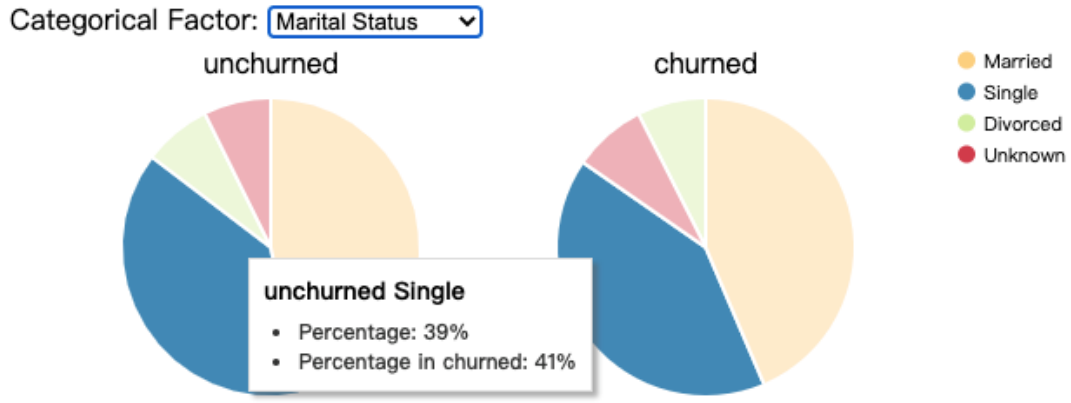
Pie chart can be effectively used to make part-to-whole judgements. Besides, the obviously different pie charts can be clearly compared, because in this case we don't care about the nuances, our focus is to find the attributes make the two pie charts significantly different, it's easy to see the most effective attributes if there are two obviously different pie charts.

Interaction:

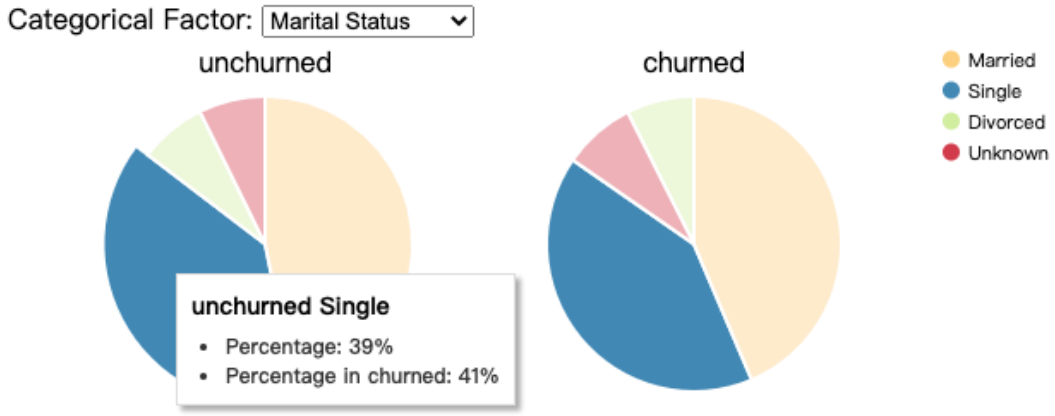
The pie chart is linked unidirectionally to the Box plot. For example: the churned box plot is selected(colored purple), then the churned pie chart will be highlighted(saturation increased).

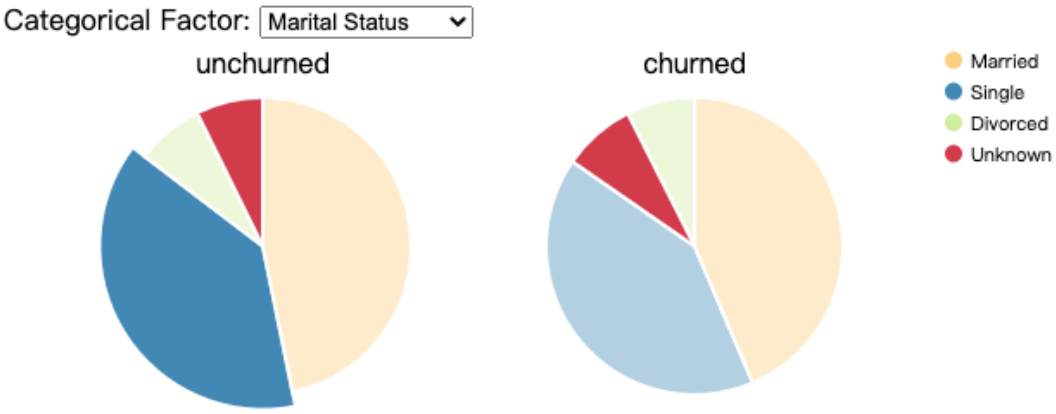


The pie is highlighted on hover. Pie chart has a tooltip when mouse over. The tooltip will show the percentage of the current hover type, and the percentage of the corresponding type in the other pie chart.



Only one type in the two pie charts could be selected at the same time: if Single in unchurned is slected, no other type could be selected at the same time. If you select Type: Single in unchurned, the tooltip will only be triggered when you hover the mouse over the pie of Single in unchurned.





There is a linked highlighting between pie charts and scatterplot: when select F in the unchurned pie chart, the pie for F will be highlighted and become larger.

Then points in the scatterplot satisfying the condition, unchurned and female, will be highlighted.

View 2 Box plot + Violin plot

Encoding:

Channel vertical position encodes attribute the range five derived quantitative attributes.

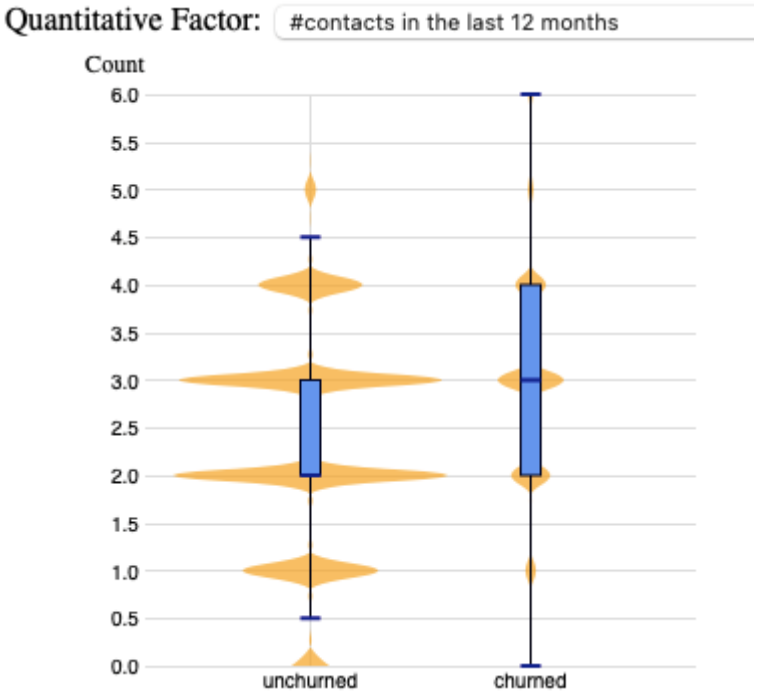
Channel horizontal spatial position encodes attribute type of customer (churned or unchurned).

Mark of type glyph encodes item boxplot.

Rationale:

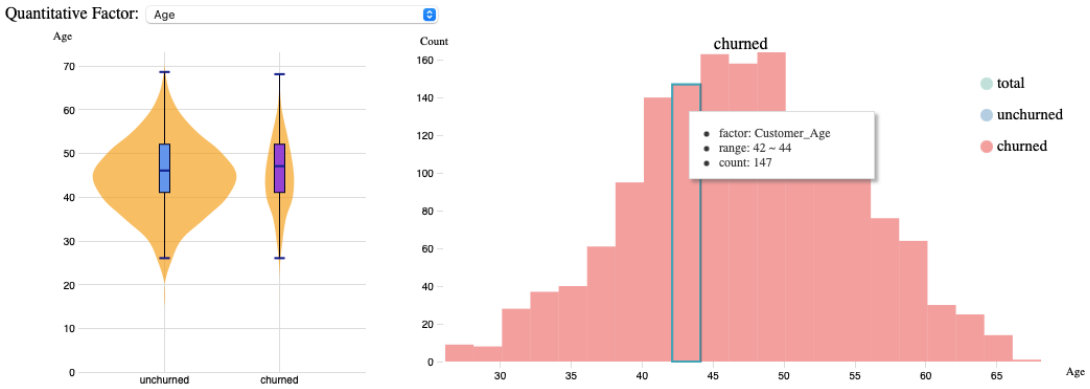
Box plot can help judge the degree and bias of data set's data dispersion, but it has a drawback that different data may have the same box plot, so we added a violin plot on the basis of it, which could more clearly reflect the dispersion density in each interval segment.

For example, in the figure below, we can see the distribution of the number of contacts segments, with 2 or 3 contacts being the most common.



Interaction:

Box plot is linked unidirectionally to Histogram: the histogram will be filtered according to the type of box selected in the box plot. If the box for the churned customer is selected, the histogram will only show the distribution of churned customer.



View 3 histogram

Encoding:

Channel vertical position encodes attribute the count of each bar.

Channel horizontal position encodes attribute derived key attribute (bin).

Mark of type line encodes item derived quantitative attribute (item count per bin).

Rationale:

The histogram is more clear for users to discover the distribution with quantitative axes, especially for those who are not familiar with the violin plot.

Interaction:

The histogram view doesn’t have interaction so far.

View 4 Scatterplot

Encoding:

Channel vertical position encodes total transaction amount.

Channel horizontal position encodes total transaction count.

Mark of type point encodes item customer.

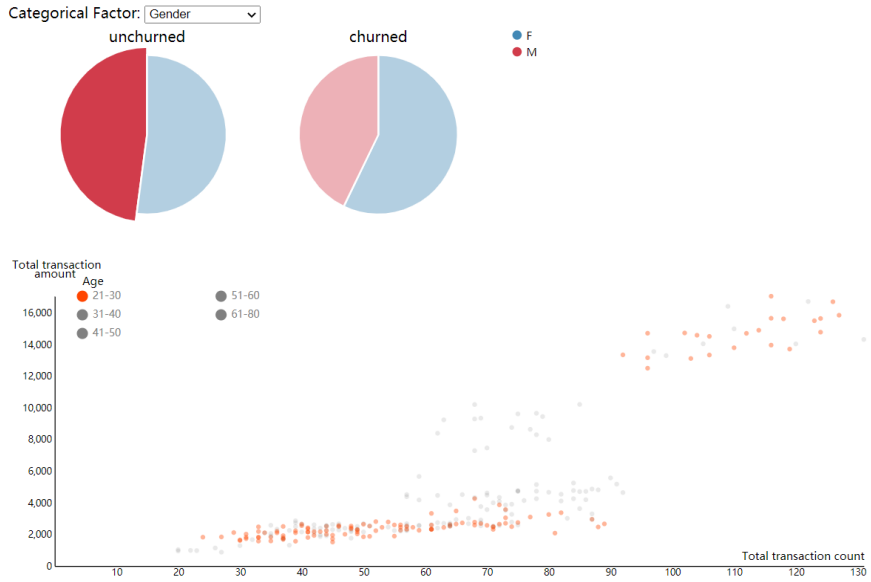
Rationale:

Scatter plot is more clear for users to find the correlation between quantitative factors: total transaction amount and total transaction count. It's also good to visualize hundreds of items at one time. Also, it shows the age distribution of the customers.

Interaction:

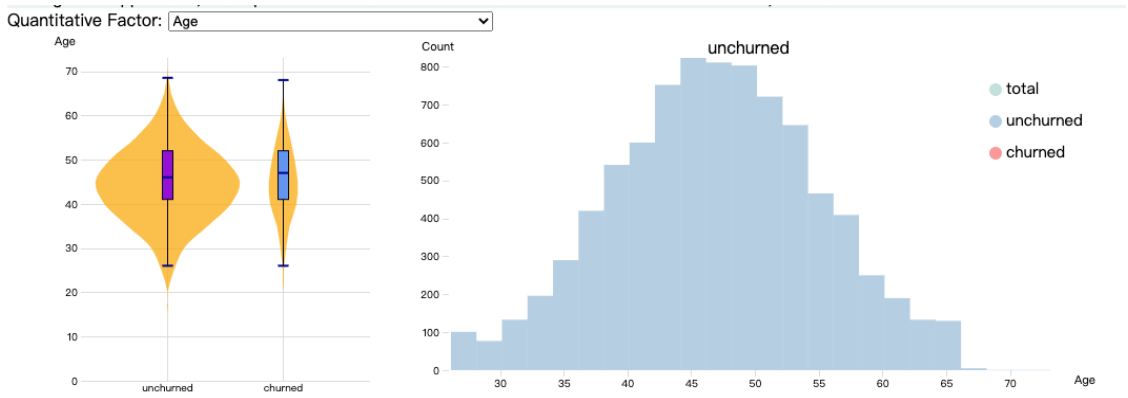
The legend in the scatterplot shows the categories of the six age levels. The legend serves as an interactive filter for the age levels. After the user clicks on an age level in the legend, it becomes active and all other categories are hidden in the chart. So far users should be able to select a single age level at a time.

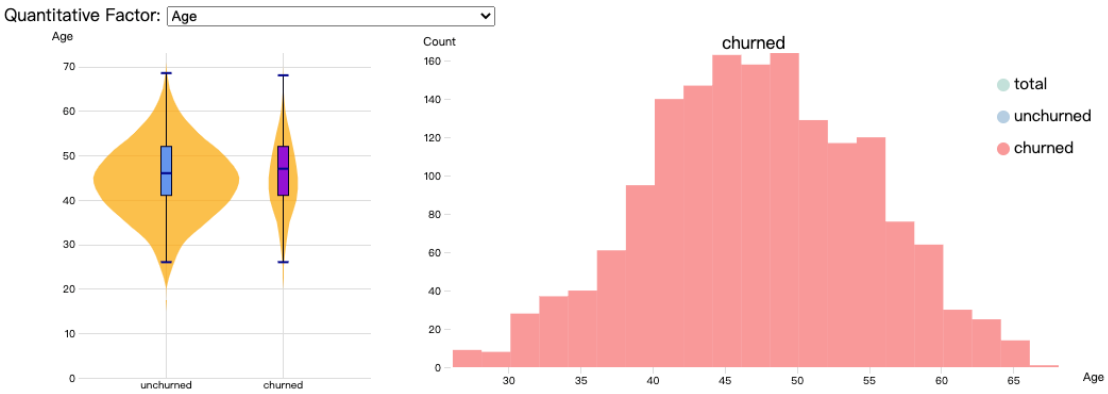
This scatterplot view is linked highlighting with pie chart as previously mentioned, that is the points of a selected factor in boxplot will be highlighted in the scatterplot. Changing the factor in pie chart will update the scatter plot as well.



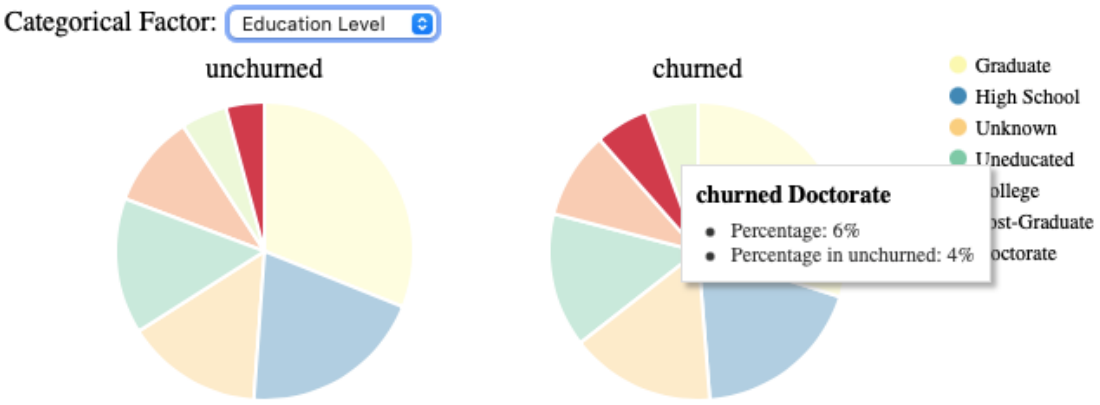
A whole walking-through story

The bank manager Joyce wants to identify which factors are majorly causing the gradual loss of customers. She starts to find out the quantitative factor, when she selects the age factor in the box plot, the distributions of churned and unchurned customer are almost the same across age, so the age factor is excluded in her analysis.

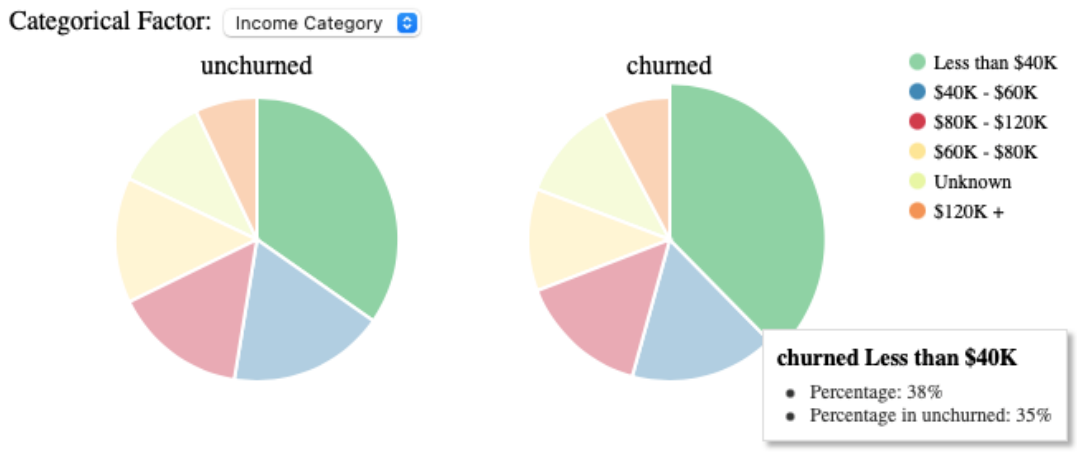




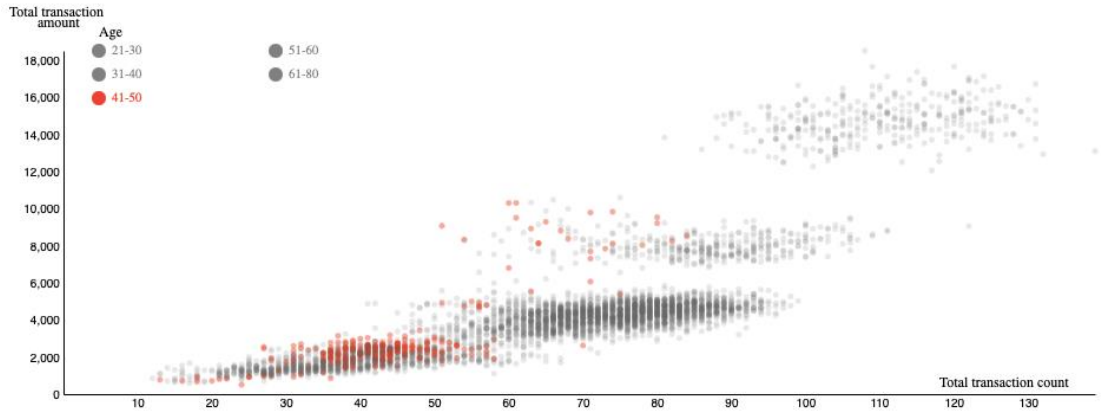
Then she tries to figure out the influential categorical factor, in the pie chart, she selects the Education level and realizes that the customers who have the Doctorate degree are more likely to churn, but these group of people are the minority of the customer.

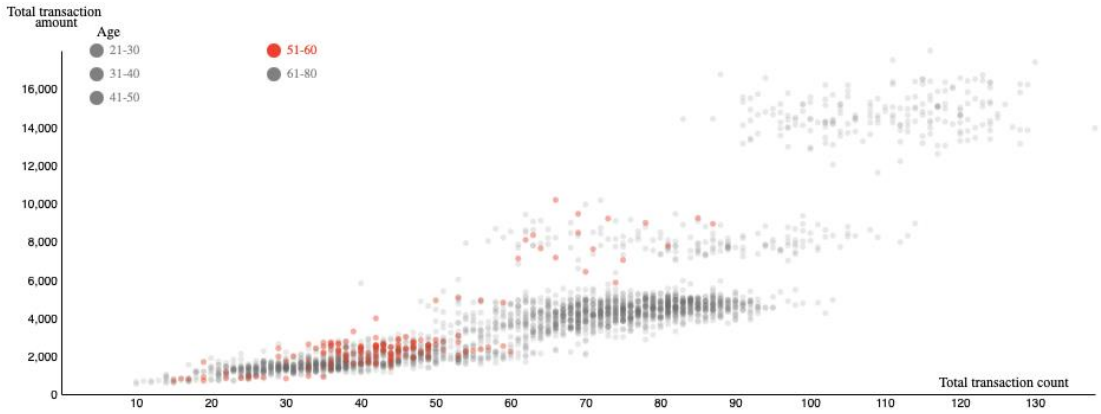


Then she selects the Income category and realizes that the customers whose income less than $40K are more likely to churn and these group of people are the majority of the customer.



Besides, in the scatterplot, she filters some age levels of churned customer respectively, then realizes that the total transaction amount and total transaction count are almost directly proportional, and the low-income population(< $40K) highlighted in the scatterplot always has a low total transaction amount with total transaction count. Therefore, this would inspire her to launch some businesses and services aimed at low-income people.





5. Credits

- https://www.d3-graph-gallery.com/graph/histogram_basic.html

- The histogram view in this project built upon this resource to set the parameters for the histogram and apply this function to data to get the bins, then append the bar rectangles to the svg elementminor tweaks. On this basis and beyond thie resource, the legend, tooltip, and interaction with boxplot were added.

- https://bl.ocks.org/rjurney/e04ceddae2e8f85cf3afe4681dac1d74

- The box plot view in this project built upon this resource to prepare the data for drawing boxplot (local min, max and box quantiles). But the implementation

in the reference link doesn't consider the 1.5 IQR. In our implementation, we consider the 1.5 IQR to delete the outliers in the boxplot (There're some outliers in some subset of data). We also referred the way of drawing the vertical lines, horizontal lines and boxes of the box plot in this resouce. On this basis and beyond thie resource, the interaction with histogram and pie chart were added.

- https://www.d3-graph-gallery.com/graph/violin_basicHist.html

- The violin plot is built upon this resource. d3.histogram is used to generate the bins needed for the violin plot. When rendering the violin plot for churned customers, we found that since the proportion of churned cursomers is small, if one bandwidth is used as the x-scale, it's difficult to obser the distribution of the violin plot. We adjust the width of the violin plot to reflect the distribution more clearly. The x-scale range for the violin plot is [0, 2.5 * bandwidth].

- https://www.d3-graph-gallery.com/graph/pie_changeData.html

- The pie charts are built upon this resource. d3.pie is used to generate the bins needed for the pie chart. Then we add legends, tooltips, interaction with mouse event to the pie charts.

- https://stackoverflow.com/questions/13669239/remove-end-ticks-from-d3-js-axis

- We used this resourse to remove the end tick in the scatter plot. Other than that, this scatter plot has a similar idea as the scatter plot we implemented in P2. The way we implemented legends is similar as the one in P1.

6. Reflection

At the beginning, the innovative veiw we designed was beeswarm plot + box plot to show the overall and detailed distribution. But we found that because we have too many data points, a chart like beeswarm that shows all individual data points is not suitibable for our dataset. We need an aggreated graph to show the detailed distribution. So we change our innovative veiw to violin plot + box plot.

Most elements in our visualizations are implemented as described in the initial proposal. Our visualization goal is unchanged that we aim to visually observe the similarities and differences between existing customers and lost

customers, but we changed the innovative view from combining two histograms of different distributions to the combination of boxplot and violin plot to show the overview visualization. Our new innovative view will display the highly summarized statistics (boxplot) and detailed distribution (violin plot) simultaneously. Although the violin plot shows the distribution, we still keep the histogram so that people who are not familiar with the violin plot can observe the distribution easily. So our technical goals have changed, in the initial proposal, there are two separate group that pie chart links with scattedplot and boxplot links with histogram, but then we added the violin plot andthe interaction logic between pie chart and boxplot. We have all fulfilled our initial proposal for the technical possibilities in D3. In addition, to make our interface as self-documenting as possible, we added a meaningful description and did more hovers than purposed to enrich the explanatory aspects of this project.

If we were to make the project again from scratch, we would think about the overview-detail views at the beginning phase,because it's a effective way to sharing the subset of data and different encoding between views, especially in the case that we hope to know the difference between churned and active customers, and the characteristics of the specific type of customers. We may consider adding scented widgets because they are better cues for information foraging.

7. Project Management & Team Assessment

Status update

J = Jiaqi, H = Harry, S = Stella

Milestone	Estimate Date	Actual Date	Estimate Hours	Actual Hours	Team Member
Data processing	Mar. 13	Mar. 13	5	4	HSJ
Static Box Plot	Mar. 16	Mar. 18	7	11	JS
Static Pie Chart	Mar. 17	Mar. 29	7	7	HS

Static Histogram	Mar. 18	Mar. 18	7	12	JS
Static Scatter Plot	Mar. 20	Mar. 28	8	8	H
Interactive Box Plot	Mar. 18	Mar. 28	5	6	JS
Interactive Pie Chart	Mar. 22	Apr. 10	6	5	SJ
~~**Compare box in Histogram**~~ **Violin Plot Layer**	Mar. 23	Mar. 27	20	7	S
Highlight in Scatter Plot and Pie Chart	Mar. 24	Apr. 14	5	5	HS
~~**Interaction between Box Plot and Pie Chart**~~ **Interaction between Box Plot and Histogram**	~~Mar. 25~~ Mar. 30	Mar. 30	15	15	SJ
Testing	Mar. 26	Mar. 26	5	3	S
Interactive between Box Plot, Pie Chart, and Scatterplot	Mar. 27	Apr. 14	20	10	HSJ
Add Select boxes	Mar. 29	Mar. 13	5	2	SHJ
M2 Write Up	Mar. 30	Mar. 31	6	7	JS~~H~~
Add Tooltip to Scatterplot and Pie Chart	Apr. 1	Mar. 29	5	4	JHS
Add Tooltip and hoverline to histogram	Apr. 2	Apr. 2	2	1.5	J
Add Legend in Scatter Plot	Apr. 3	Mar. 28	10	4	H

Add More Style in CSS	Apr. 5	Apr. 14	4	3	J
Add Titles and Labels	Apr. 6	Mar. 31	2	2	~~H~~J
M3 Write Up	Apr. 10	Apr.14	8	10	SHJ

Contributions breakdown

The progress of the work so far is as follows:

Stella:

- Mainly leaded view designs. Redesign the innovative view, and build the violin plot.
- Built the pie chart view based on Harry's general implementation for the pie chart. Add tooltip and mouse event for hovering and clicking in the pie chart view.
- Built interaction between box plot and pie charts. Built select box.
- Implemented pre-processing code for interaction between pie charts and scatterplot.
- Added axis titles and legend title in the scatterplot.
- Built the box plot and histogram together with Jiaqi.
- Changed the webpage layout and CSS.
- Wrote the writeup: Proposal, M2, M3. Wrote the high-level documentation in README.md

Harry:

- Built the interactable scatterplot.
- Added tooltip and legends in the scatterplot.
- Implemented pie chart with a select box (filter).
- Wrote the writeup: sketches and work breakdown in Proposal, scatterplot part revisal in M3.
- Built the interaction between pie chart and scatterplot.

Jiaqi:

- Mainly leaded whole writeups: Proposal, M2, M3.
- Built a select box in the boxplot.
- Built the static box plot and static histogram together with Stella.

- Added axis labels in the boxplot and histogram.
- Added mouse event for hovering and clicking in the boxplot view.
- Added the interaction between the boxplot view and histogram view.
- Added title, description and background on the webpage.
- Added tooltip in the histogram to show the width range and detailed information for each bar.
- Adjusted the final page layout and CSS style.