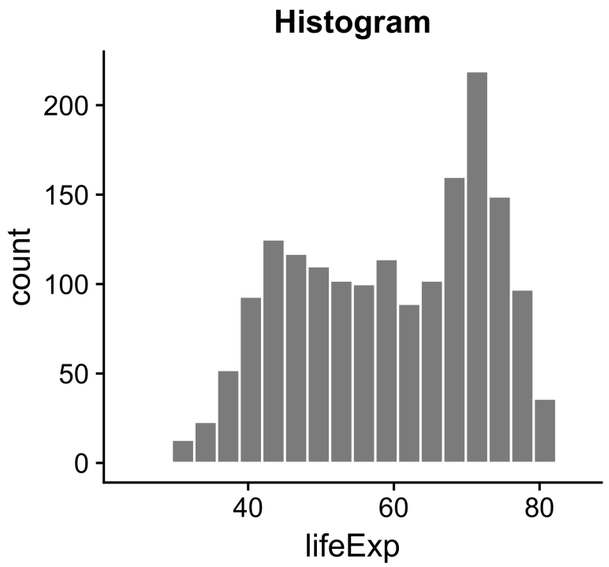
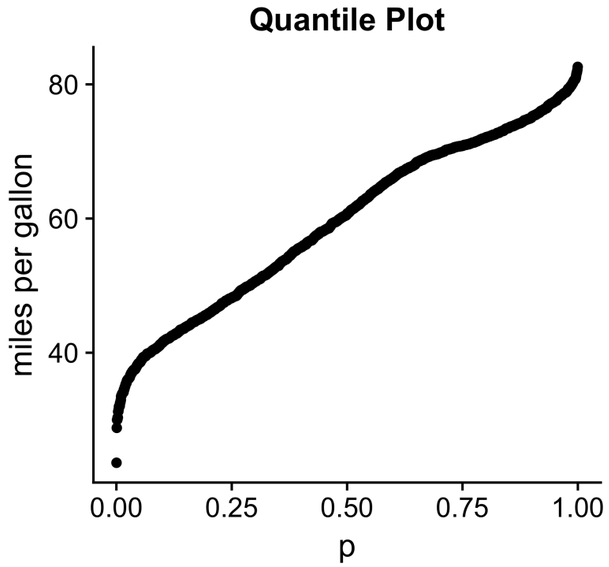
VISUALIZING BIG DATA WITH TRELLISCOPE

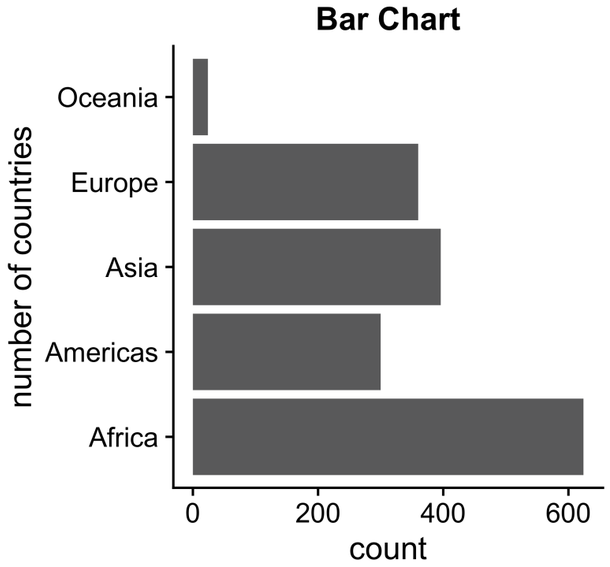
Introduction

Ryan Hafen
Author, TrelliscopeJS

Overview


Histogram
count
200
150
100
50
0
40
60
80
lifeExp


Quantile Plot
miles per gallon
80
60
40
0.00
0.25
0.50
0.75
1.00
p


Bar Chart
number of countries
Oceania
Europe
Asia
Americas
Africa
0
200
400
600
count

Summaries of One Variable

- **Continuous variables**
- **Categorical variables**
- **Temporal variables**

Gapminder Data

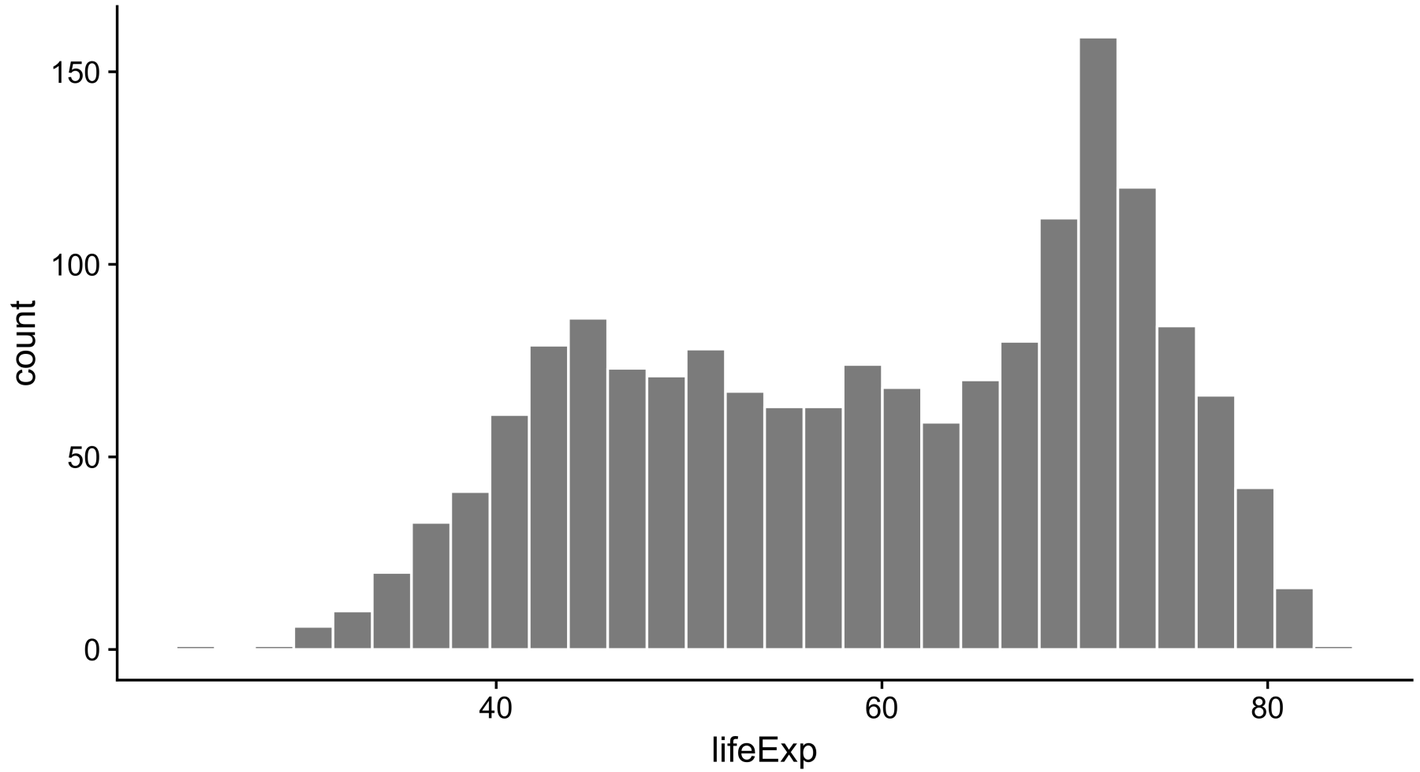
```
library(gapminder)

gapminder
```

```
# A tibble: 1,704 x 6
   country     continent  year lifeExp      pop gdpPercap
   <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
 1 Afghanistan Asia       1952    28.8  8425333      779.
 2 Afghanistan Asia       1957    30.3  9240934      821.
 3 Afghanistan Asia       1962    32.0 10267083      853.
 4 Afghanistan Asia       1967    34.0 11537966      836.
 5 Afghanistan Asia       1972    36.1 13079460      740.
 6 Afghanistan Asia       1977    38.4 14880372      786.
 7 Afghanistan Asia       1982    39.9 12881816      978.
 8 Afghanistan Asia       1987    40.8 13867957      852.
 9 Afghanistan Asia       1992    41.7 16317921      649.
10 Afghanistan Asia       1997    41.8 22227415      635.
# ... with 1,694 more rows
```

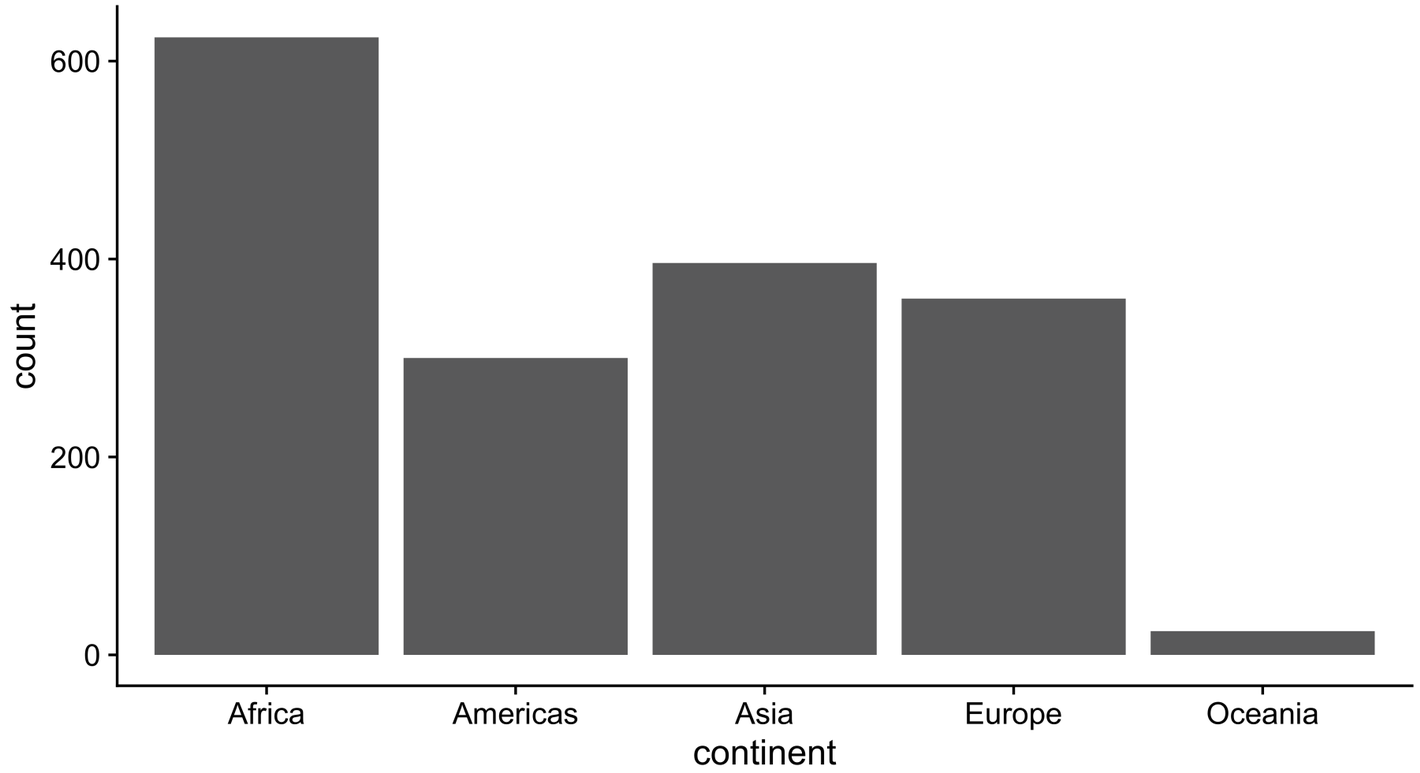
Summaries of One Variable: Continuous

```
ggplot(gapminder, aes(lifeExp)) +
  geom_histogram()
```



Summaries of One Variable: Discrete

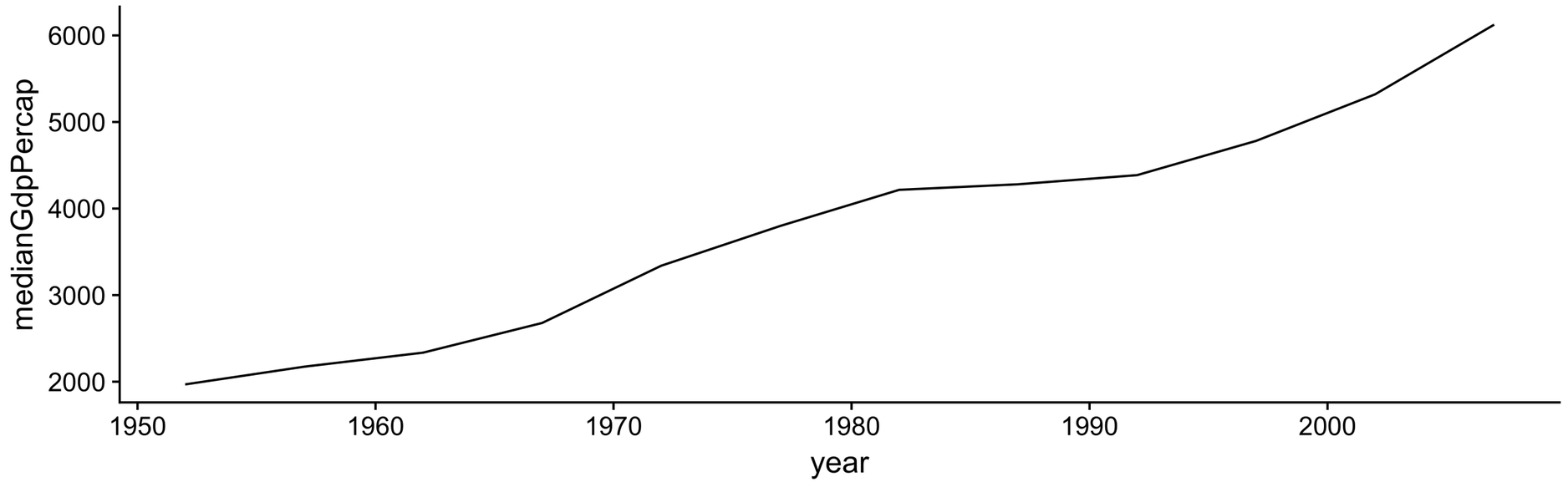
```
ggplot(gapminder, aes(continent)) +
  geom_bar()
```



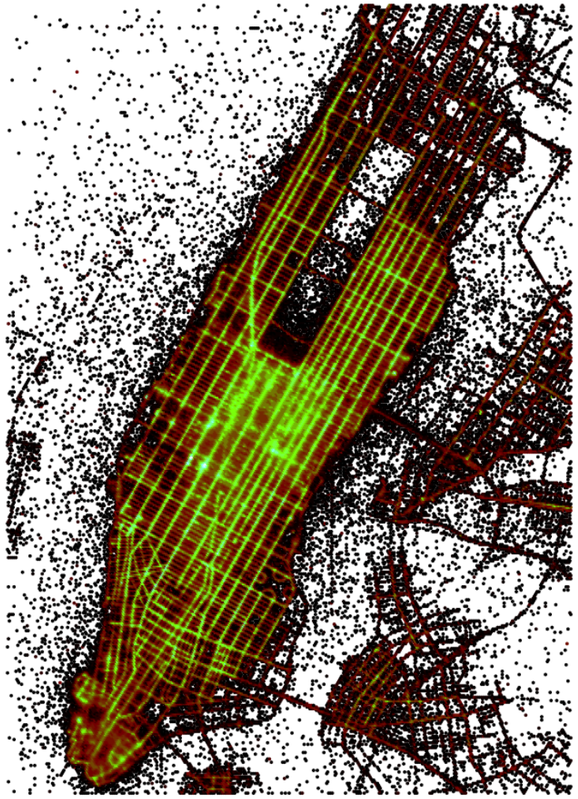
Summaries of One Variable: Temporal

```
by_year <- gapminder %>%
  group_by(year) %>%
  summarise(medianGdpPercap = median(gdpPercap, na.rm = TRUE))

ggplot(by_year, aes(year, medianGdpPercap)) +
  geom_line()
```



1 Million NYC Taxi Rides



Random sample of rides from July to December 2016

Taxi Data

```
> tx
# A tibble: 1,000,000 x 7
     pick_day pick_dow total_amount tip_amount payment_type trip_duration
       <date>   <fctr>        <dbl>      <dbl>       <fctr>         <dbl>
 1 2016-07-09      Sat        47.60      23.80         Card     26.116667
 2 2016-07-28      Thu         9.96       1.66         Card      5.866667
 3 2016-07-20      Wed         6.80       1.00         Card      4.916667
 4 2016-07-30      Sat        11.75       1.95         Card     10.350000
 5 2016-07-19      Tue         7.30       0.00         Cash      6.866667
 6 2016-07-07      Thu        12.05       2.75         Card      7.050000
 7 2016-07-29      Fri        13.80       0.00         Cash     13.700000
 8 2016-07-17      Sun        14.16       2.36         Card     13.233333
 9 2016-07-18      Mon        13.30       0.00         Cash     13.666667
10 2016-07-14      Thu        21.80       2.00         Card     29.316667
# ... with 999,990 more rows, and 1 more variables: pick_wkday <lgl>
```

VISUALIZING BIG DATA WITH TRELLISCOPE

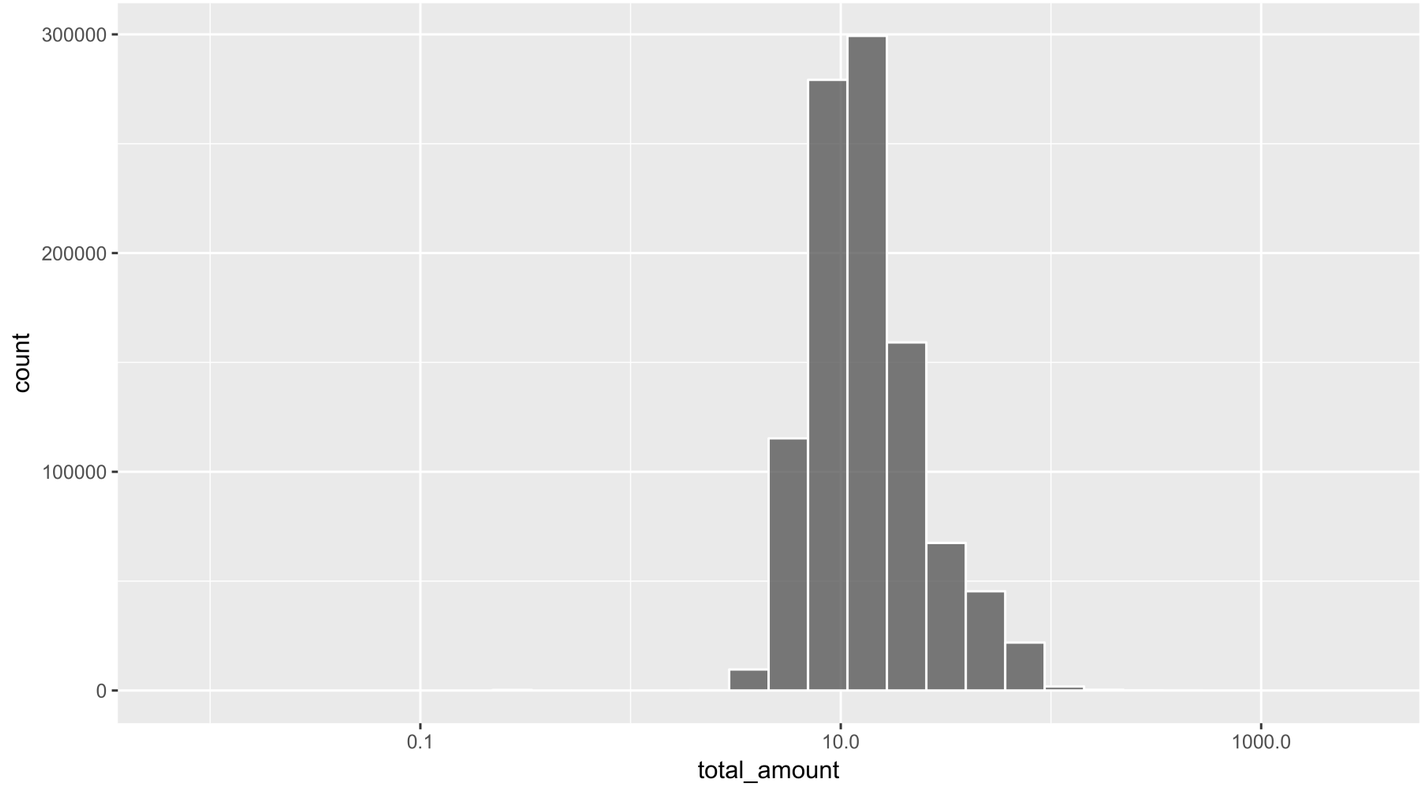
Let's practice!

VISUALIZING BIG DATA WITH TRELLISCOPE

Adding More Detail to Summaries

Ryan Hafen
Author, TrelliscopeJS

Distribution of Total Fare Amount



Adding more detail to summaries

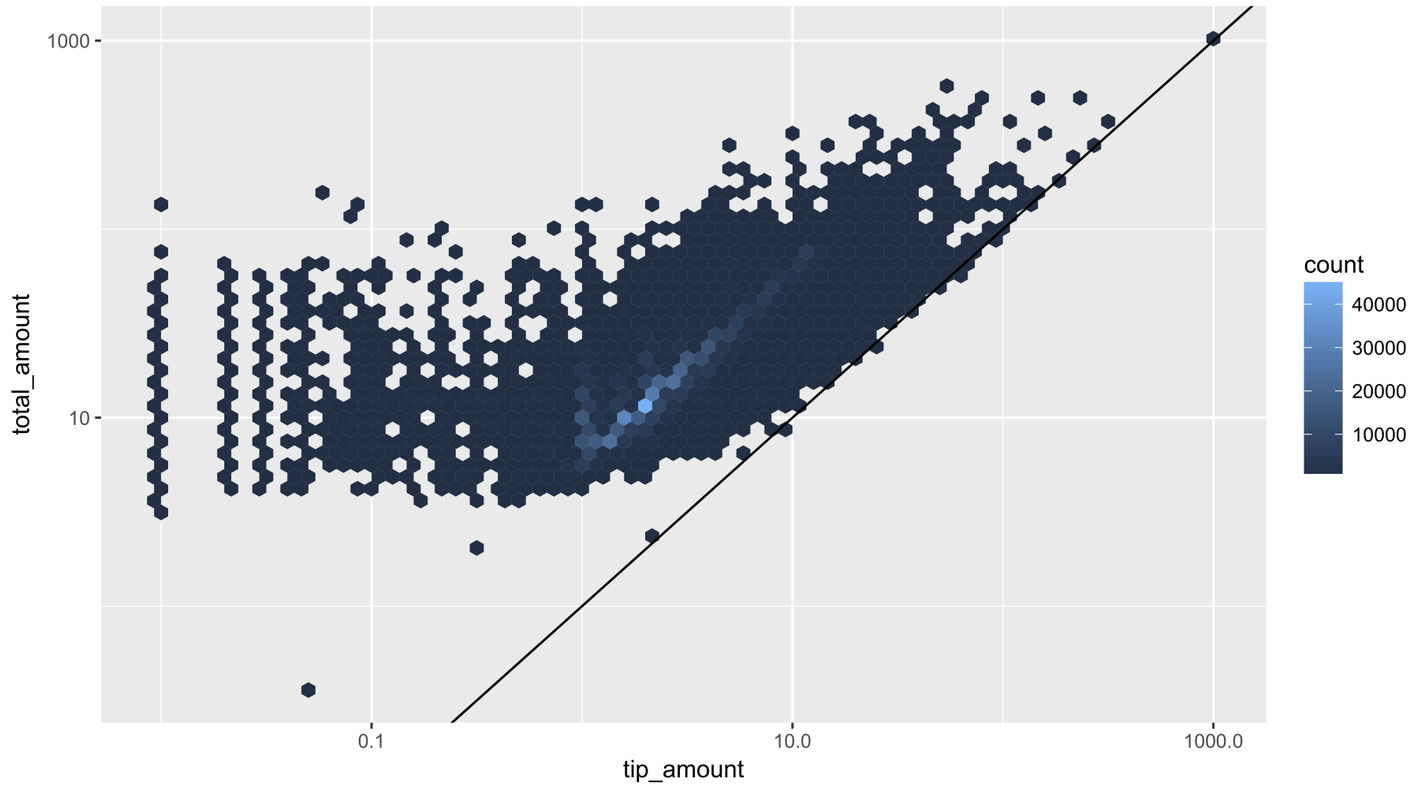
Introduce more variables into the summary computations

- **Binning** two or more continuous variables to visualize joint distribution
- **Grouping** or **faceting** summary computations by additional variables

Binning Continuous Variables using geom_hex()

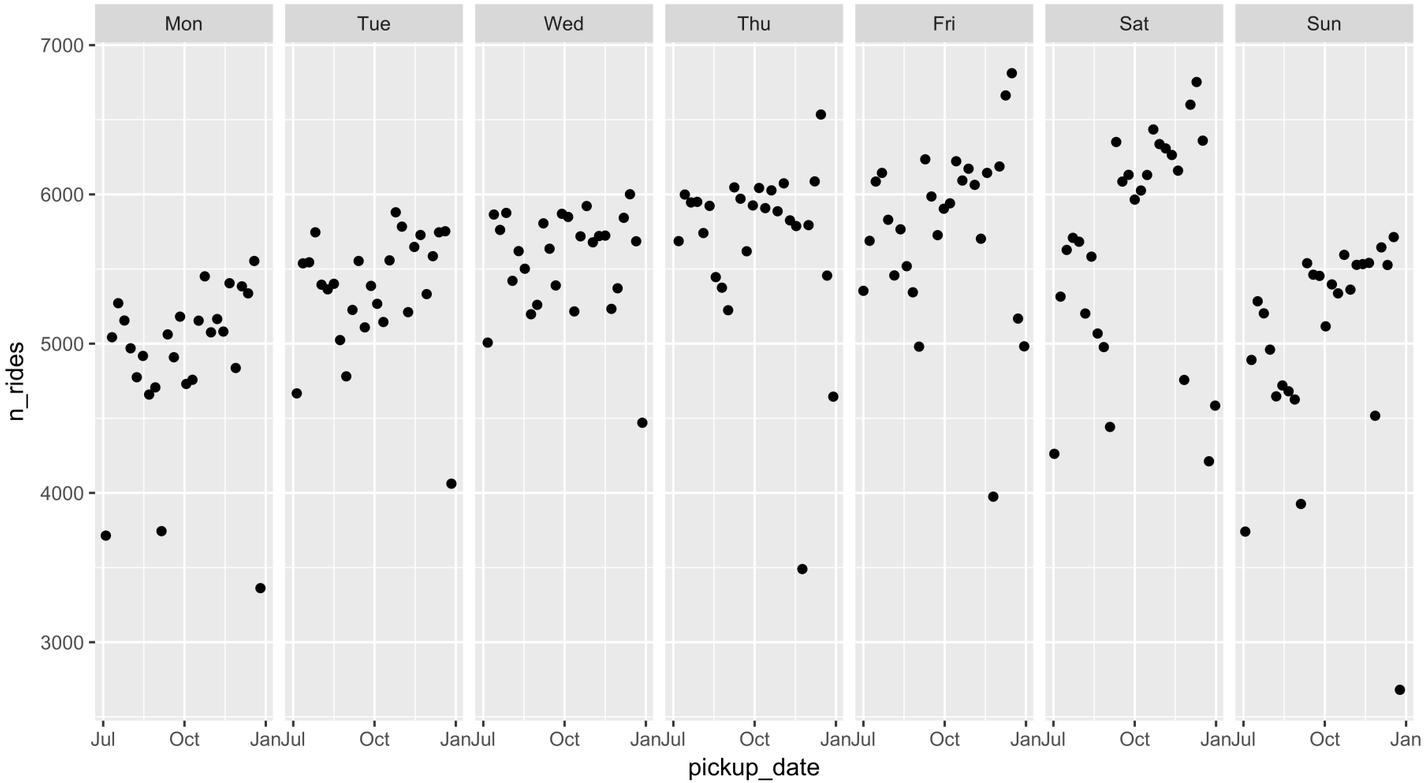
```
ggplot(tx, aes(tip_amount, total_amount)) +

  geom_hex(bins = 75) +
  scale_x_log10() + scale_y_log10() +
  geom_abline(slope = 1, intercept = 0)
```



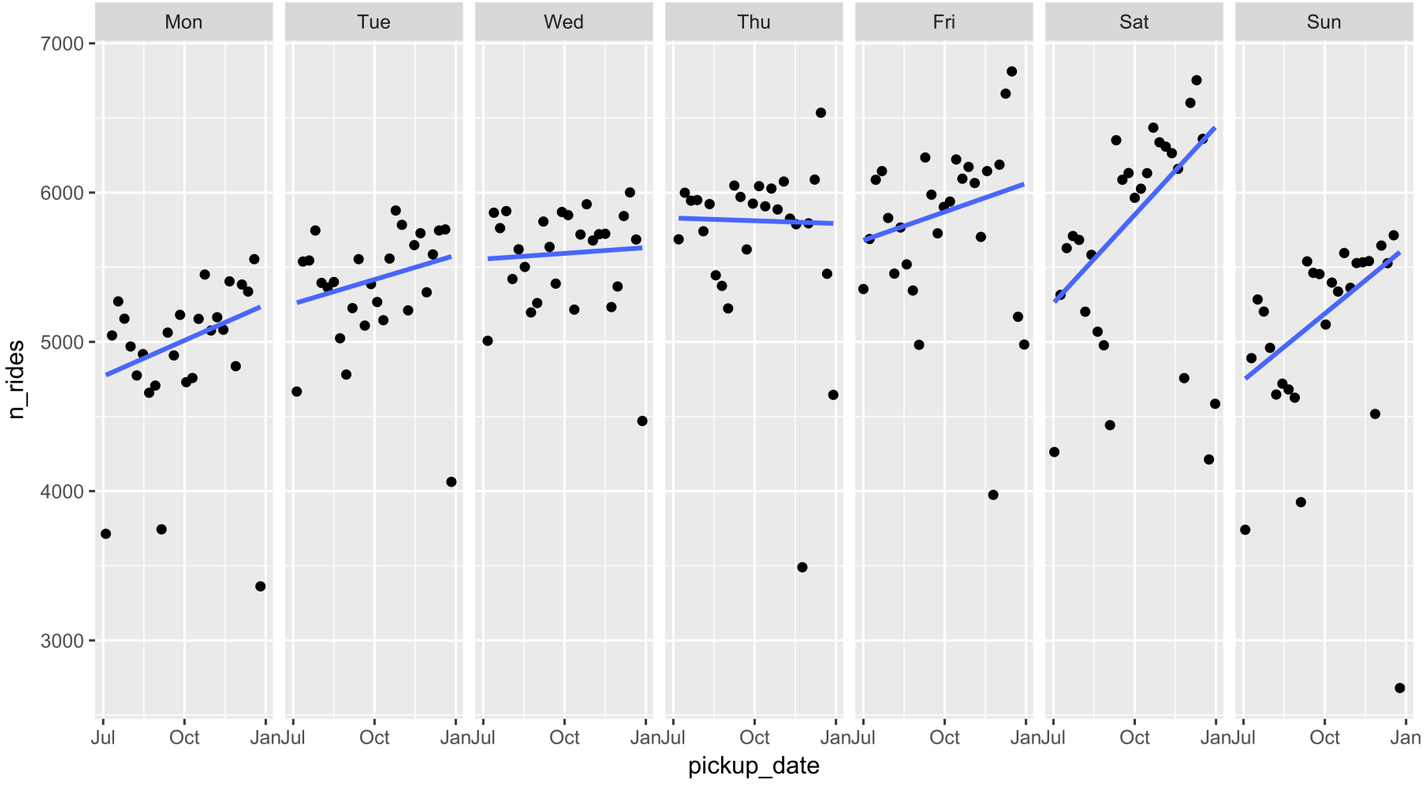
Faceting using facet_wrap()

```
ggplot(daily_count, aes(pickup_date, n_rides)) +
  geom_point() +
  facet_wrap(~ pickup_dow)
```



Faceting

```
ggplot(daily_count, aes(pickup_date, n_rides)) +
  geom_point() +
  facet_grid(~ pickup_dow) +
  geom_smooth(method = "rlm", se = FALSE)
```



VISUALIZING BIG DATA WITH TRELLISCOPE

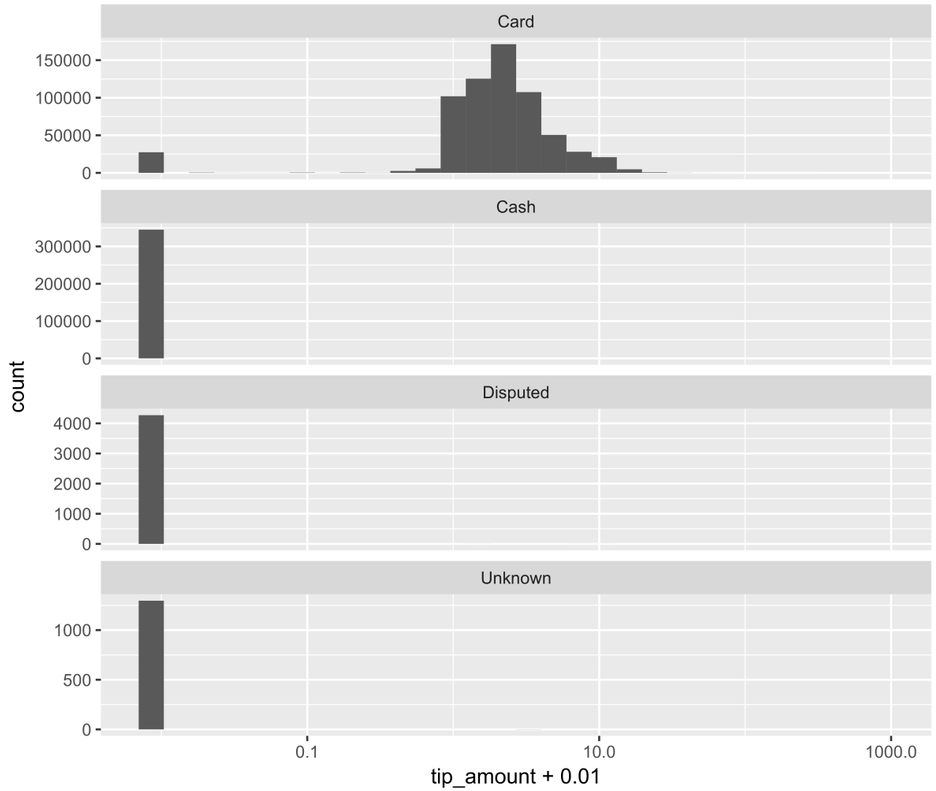
Let's practice!

VISUALIZING BIG DATA WITH TRELLISCOPE

Visualizing Subsets

Ryan Hafen
Author, TrelliscopeJS

Investigating the Tip Amount Distribution

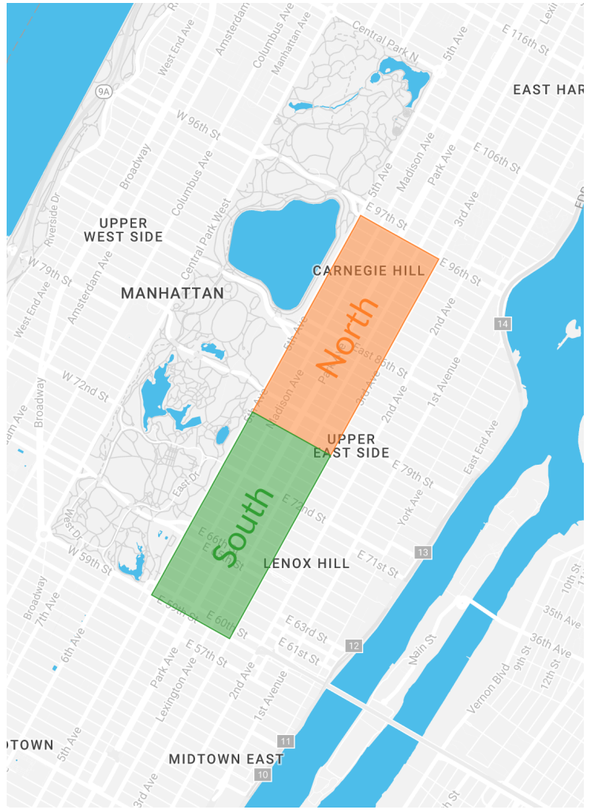


Question: Do cash payments have tips?

A Subset of the Taxi Data

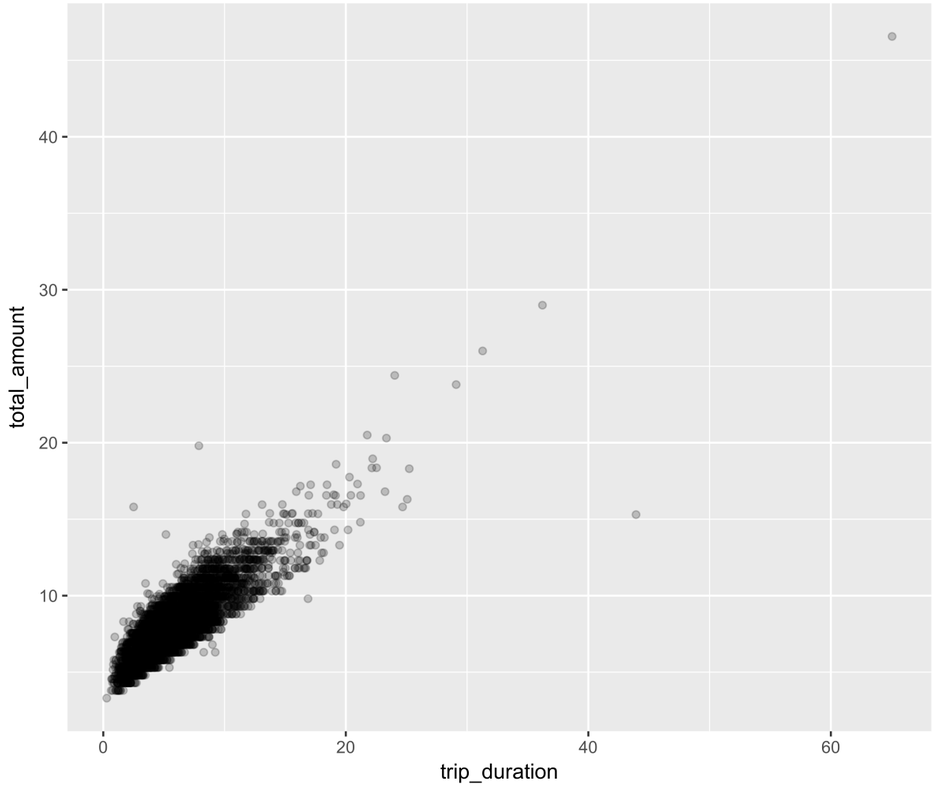
Rides of the same nature should have similar fare and tip amounts.

- **Most popular route:** Upper East Side South to the Upper East Side North of Manhattan
- Only include cash and credit transactions
- 5,187 observations



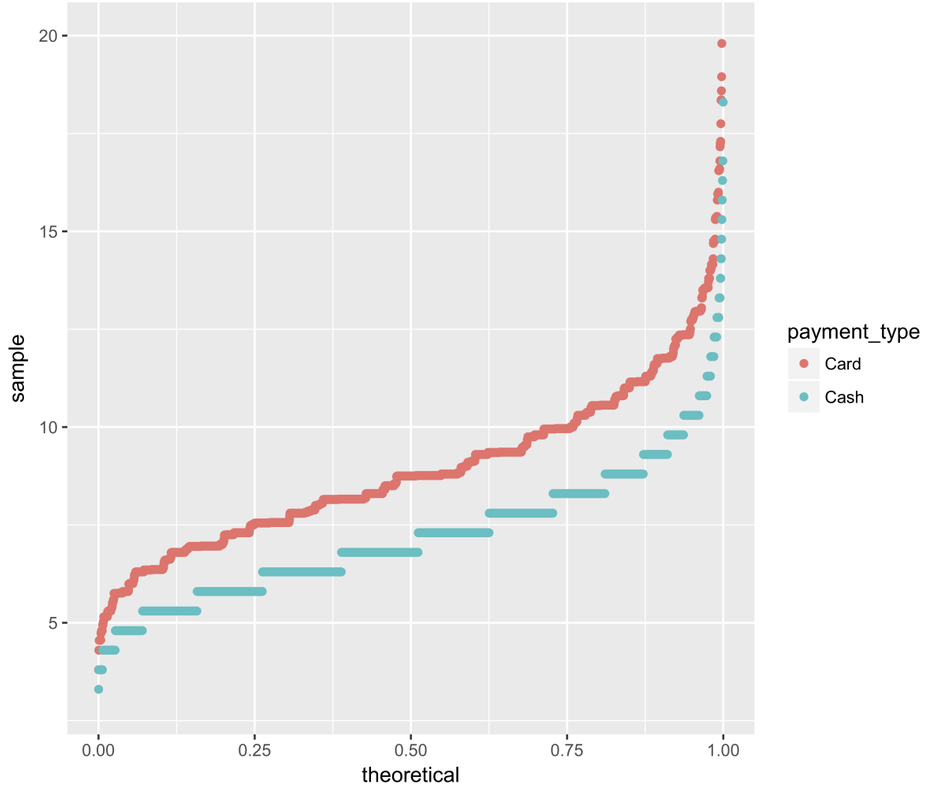
Total Fare vs. Trip Duration

```
ggplot(tx_pop, aes(trip_duration, total_amount)) +
  geom_point(alpha = 0.2)
```



Cash / Card Distribution Comparison Using a Quantile Plot

```
ggplot(tx_pop, aes(sample = total_amount, color = payment_type)) +
  geom_qq(distribution = stats::qunif) +
  ylim(c(3, 20))
```



VISUALIZING BIG DATA WITH TRELLISCOPE

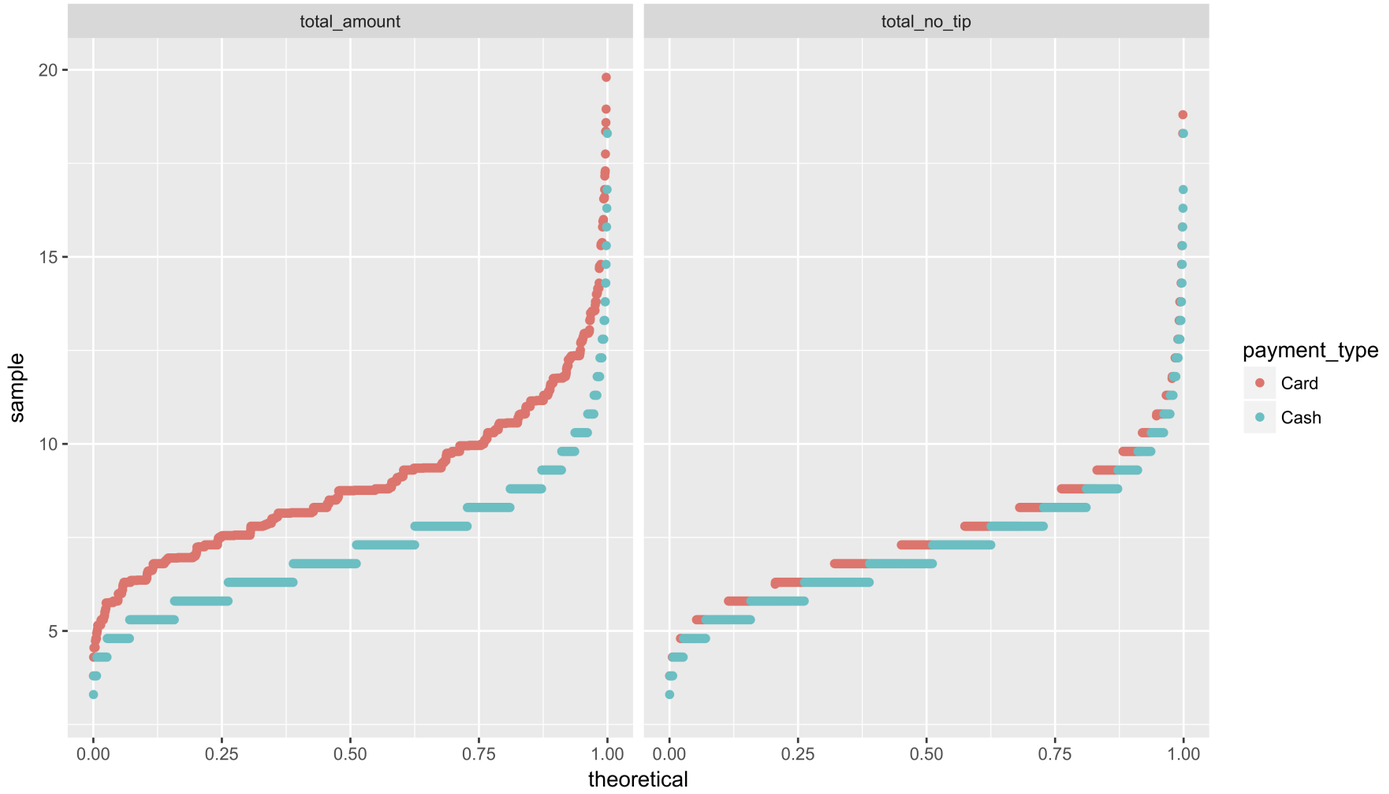
Let's practice!

VISUALIZING BIG DATA WITH TRELLISCOPE

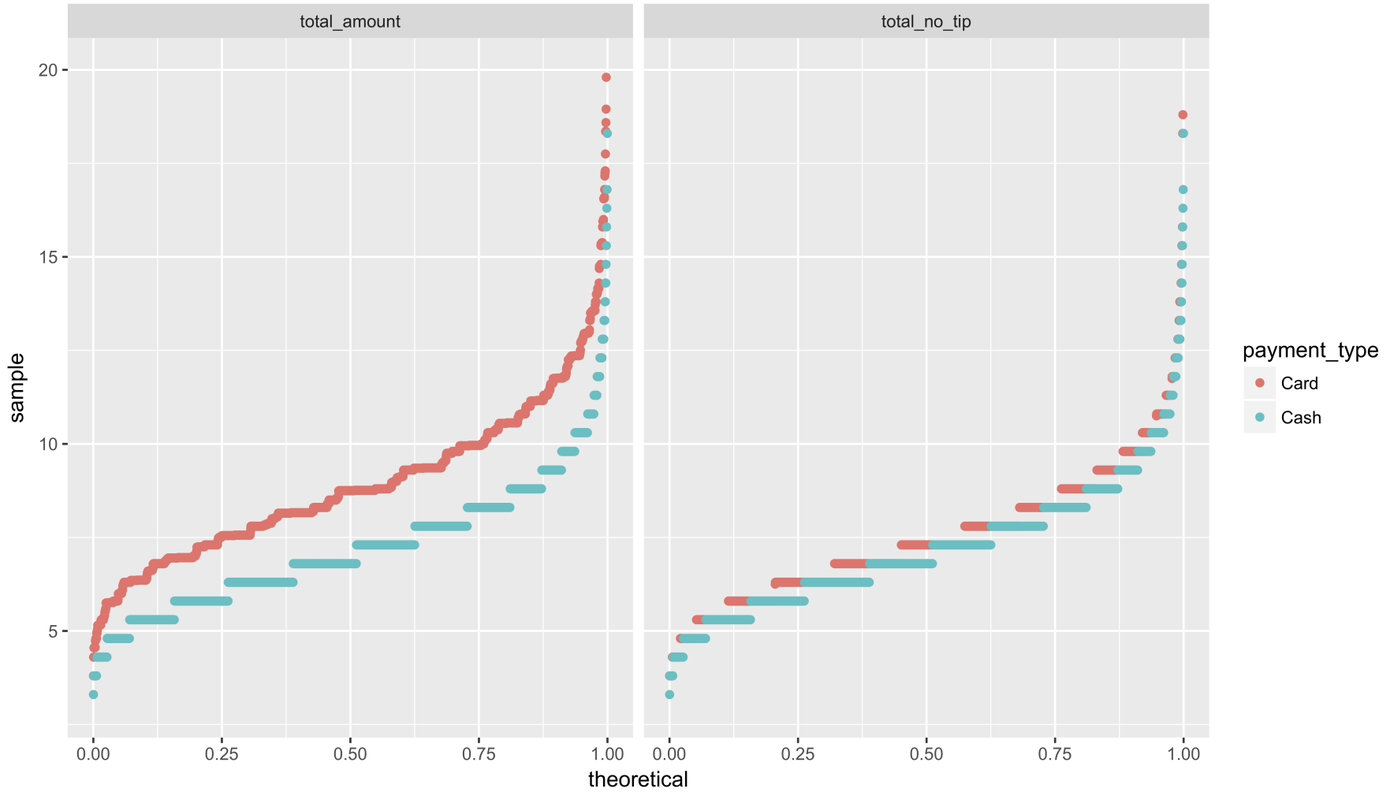
Visualizing All Subsets

Ryan Hafen
Author, TrelliscopeJS

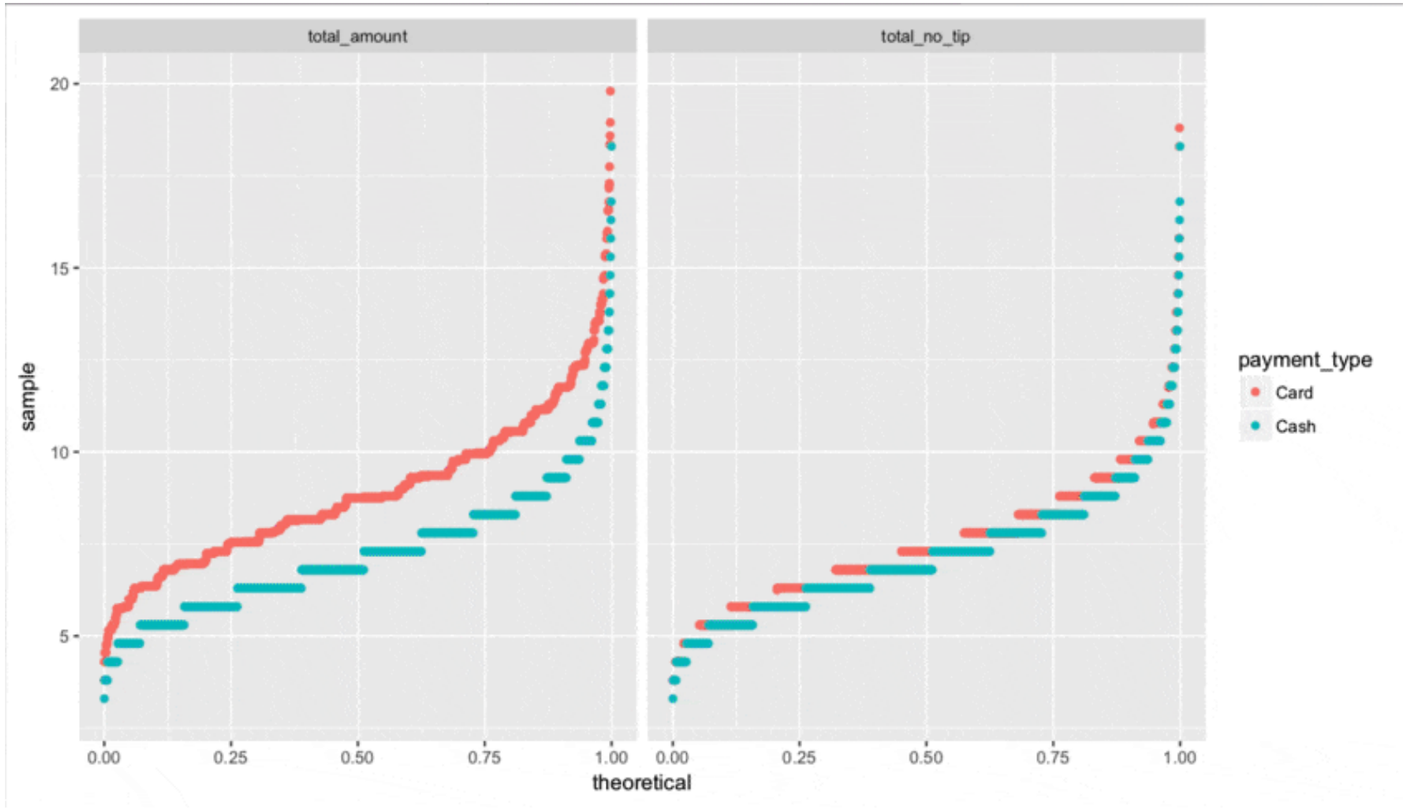
Insight From Subset Exploration



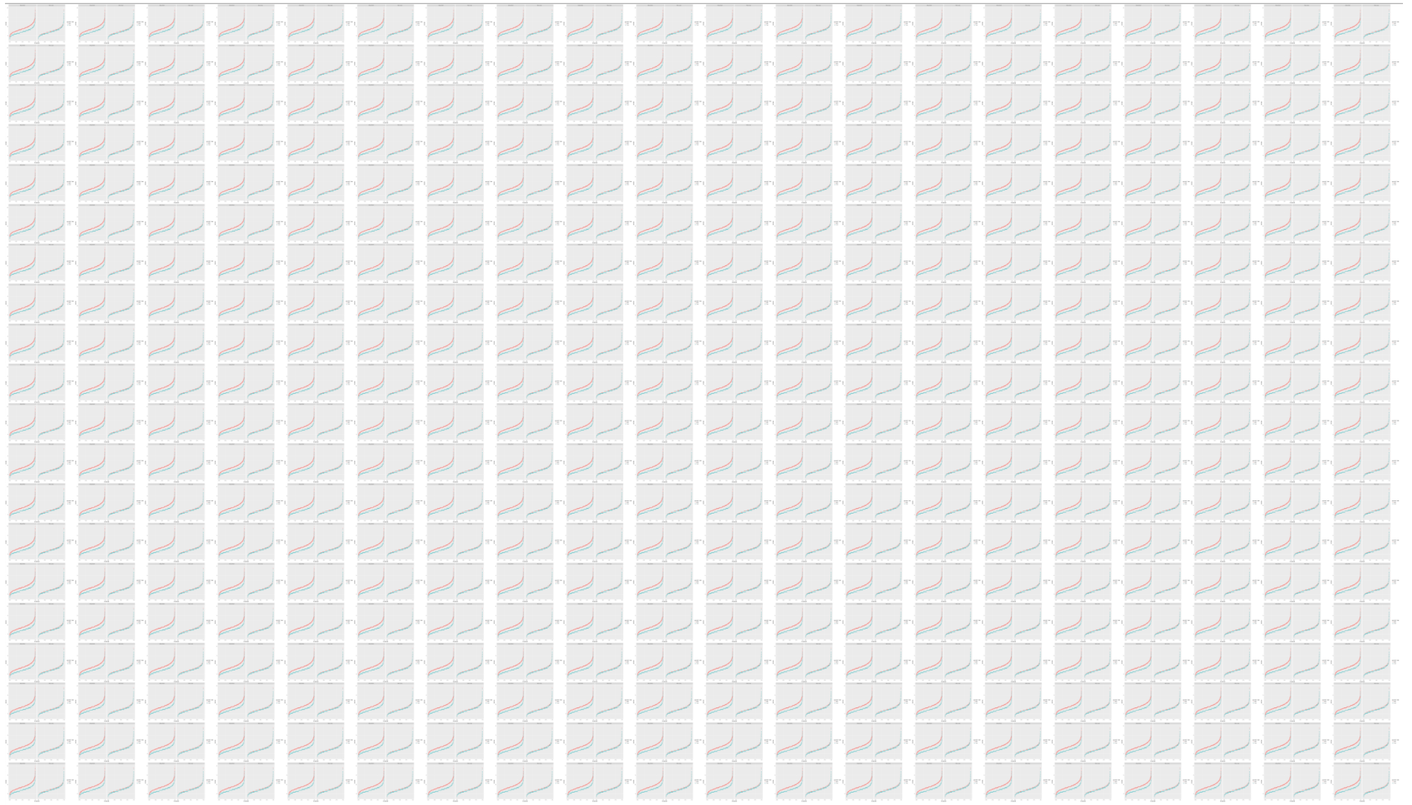
Visualizing All Subsets



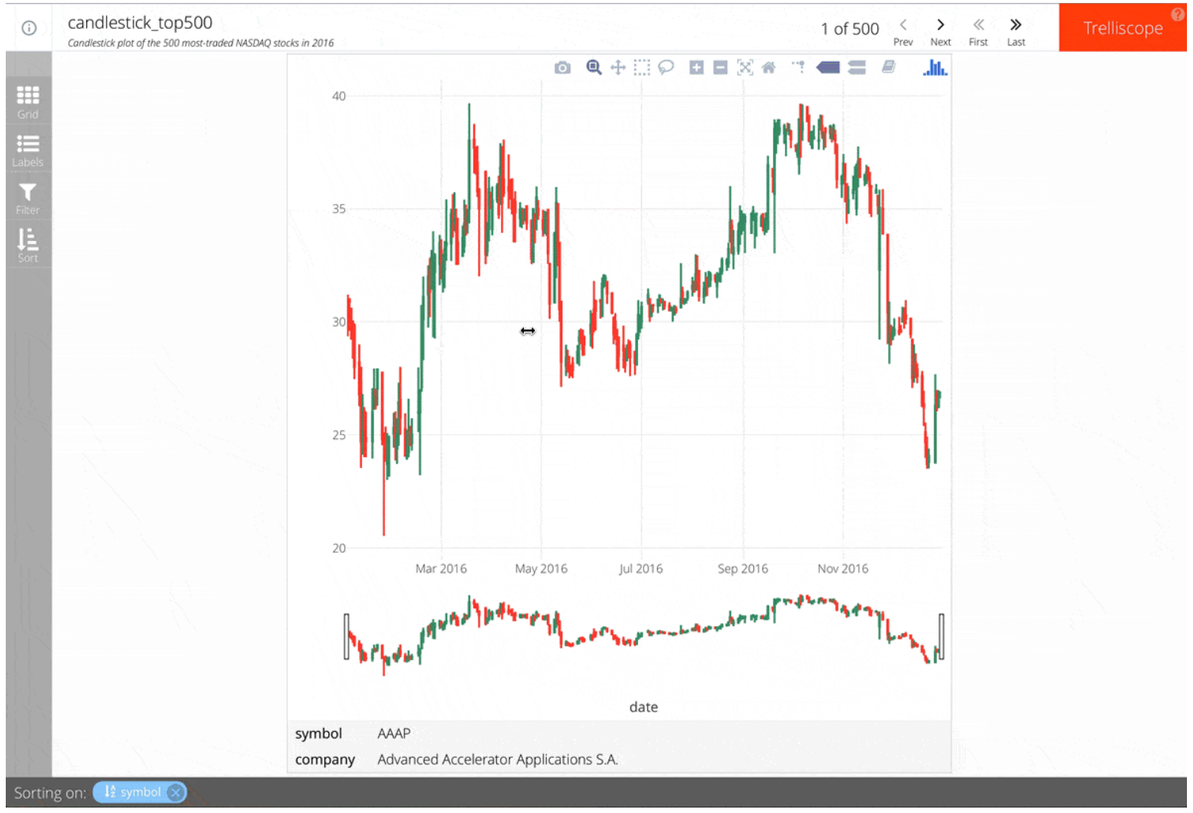
Visualizing All Subsets



Visualizing All Subsets with Trelliscope



Visualizing All Subsets with Trelliscope



VISUALIZING BIG DATA WITH TRELLISCOPE

See you in Chapter 2!