Li_PS1

Li Liu

October 10, 2019

Exploration & Computation

Q1.Obtain a dataset

The data is about people’s shopping behaviors when buying grocery and gourmet food on Amazon from May 1996 to July 2014.

There are 36978 observations and 16 variables.

```
meta<- read.csv("C:/Users/lliu9/Desktop/Amazon Reviews/meta_Grocery_and_Gourmet_Food.csv")
```

Q2.Visual technique (boxplot)

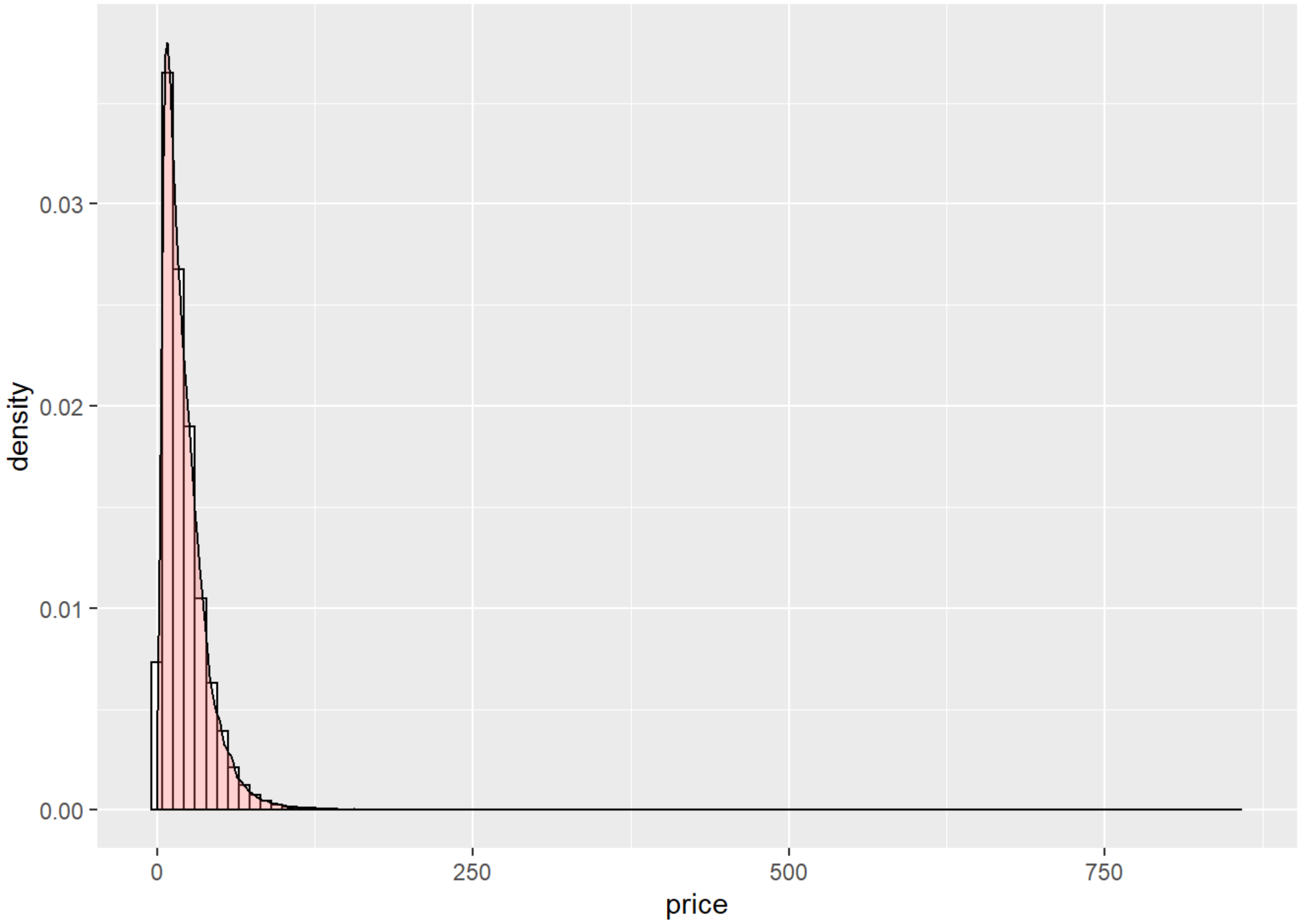
I will use the histogram and kernel density to show the distribution of grocery prices.

Q3.Visualization

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```
ggplot(meta, aes(x=price)) +
    geom_histogram(aes(y=..density..), colour="black", fill="white", bins=100)+
    geom_density(alpha=.3, fill="#FF6666")
```



Description: Most of the prices seems to be below 100. The distribution has a long tail.

Q4. Central tendency and variation

```
summary(meta$price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.01    9.03   16.99   22.03   29.00  854.80
```

```
sd(meta$price)
```

```
## [1] 20.7638
```

Q5. Numeric output

a. It reveals that prices differ a lot in this data. The data is right skewed (mean>median). There are some transactions that have very high prices.
b. It is important to further study the effect of other factors (brand, reviews, etc) on prices.
c. We could segment the products into several groups. We can also look at how people make decisions differently for products with low and high prices.

Critical Thinking

Q1. Visual vs Numeric

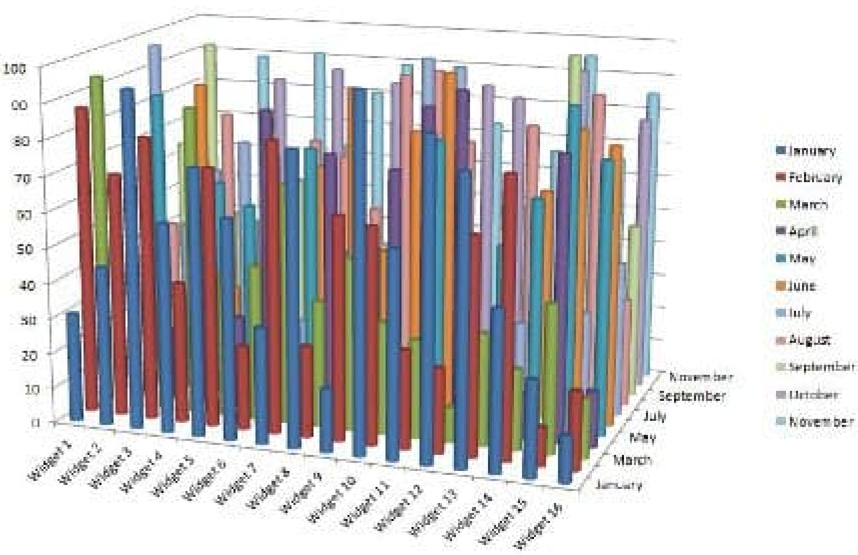
Visual EDA could present the information in a more intuitive way, especially for spatial data, social networks, text, etc.

Numeric EDA reveals the important statistics for a brief look. It is helpful for detecting outliers and anomalies when data is large.

Q2.Bad visualizations

Example 1.

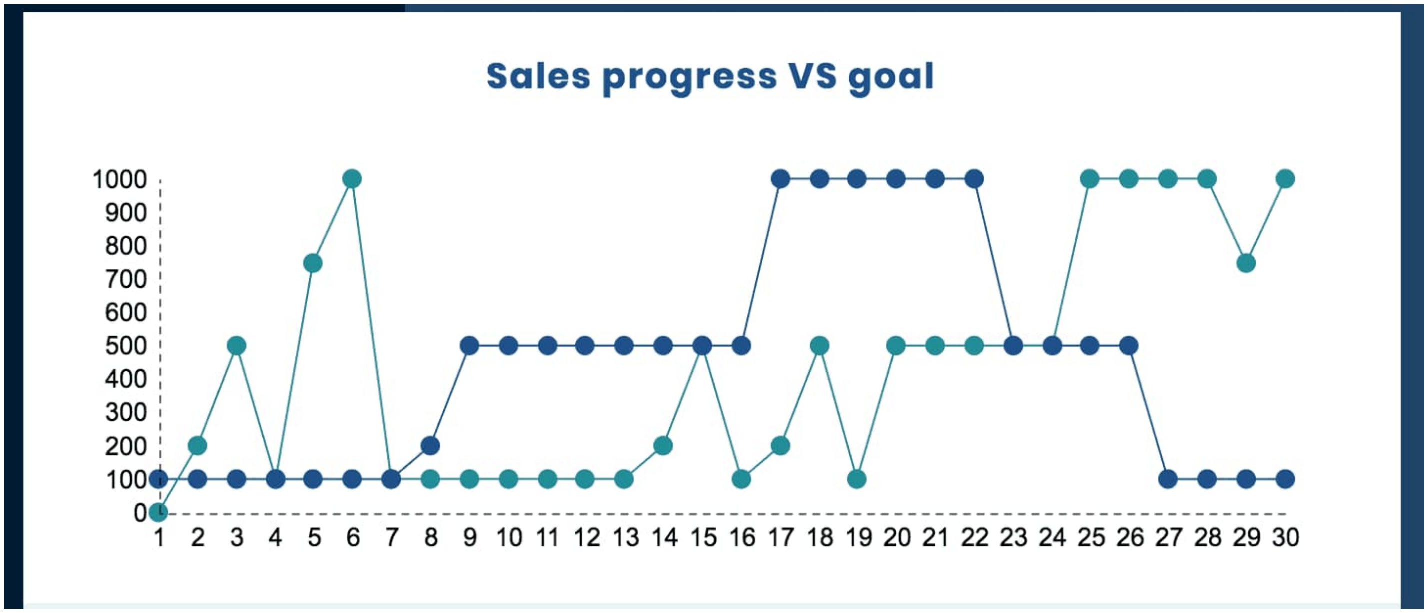
Source: https://blog.hubspot.com/sales/bad-data (https://blog.hubspot.com/sales/bad-data)



Bad visualization!

This visualization tries to show the sales distribution of 16 widgets in different months. However, viewers cannot find useful information and might get confused from this plot as there are too many bars.

Example 2.



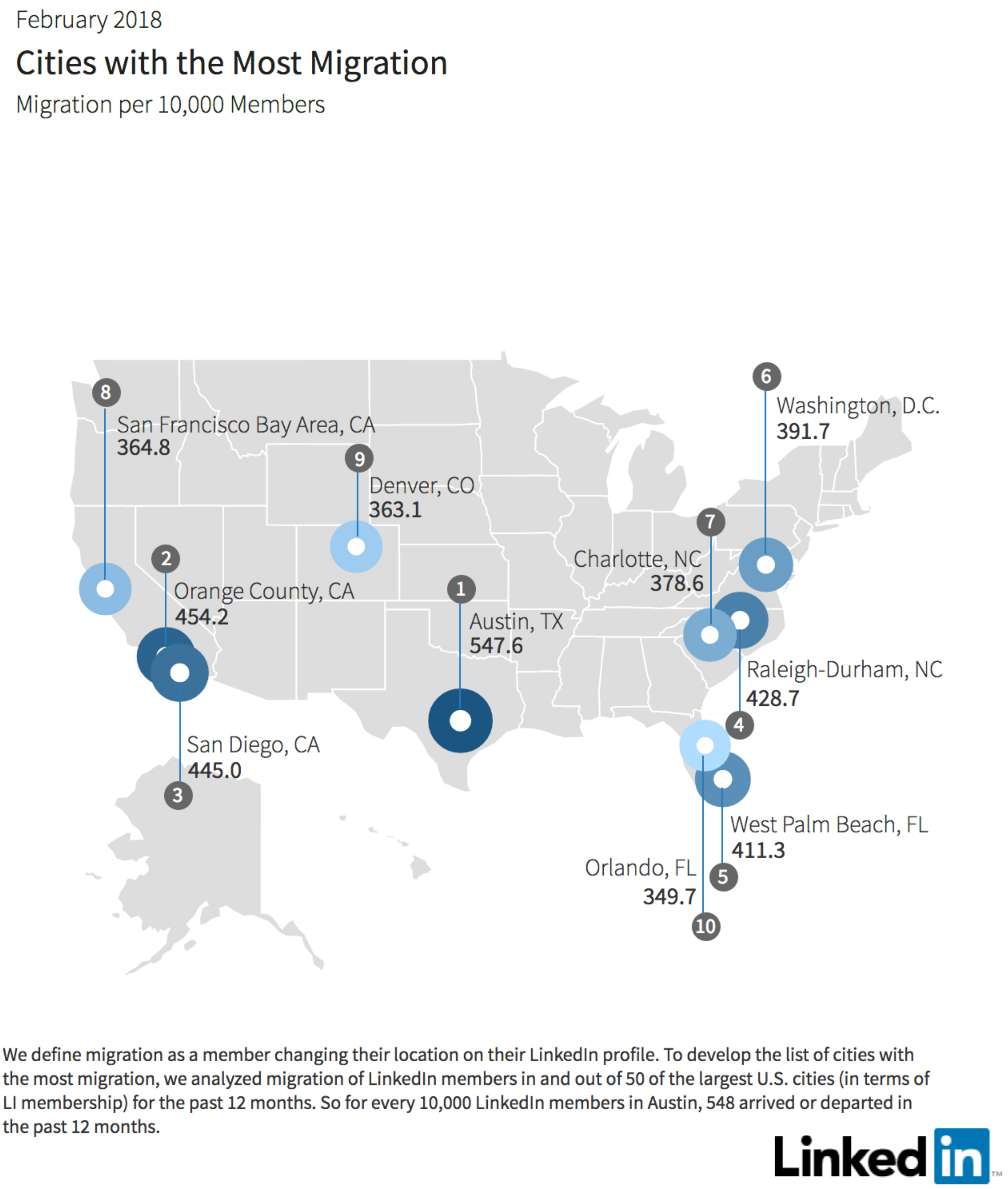
Bad visualization!

This visualization wants to show the gap between sales progress and goal. However, it doesn’t account for the seasonality or month-to-month holidays. The units for the x and y labels are also missing. So managers cannot get actionable insight from this.

Q3. Good visualizations

Example 1.

Source: https://www.columnfivemedia.com/data-storytelling-brands-data-visualization (https://www.columnfivemedia.com/data-storytelling-brands-data-visualization)



February 2018

Cities with the Most Migration

Migration per 10,000 Members

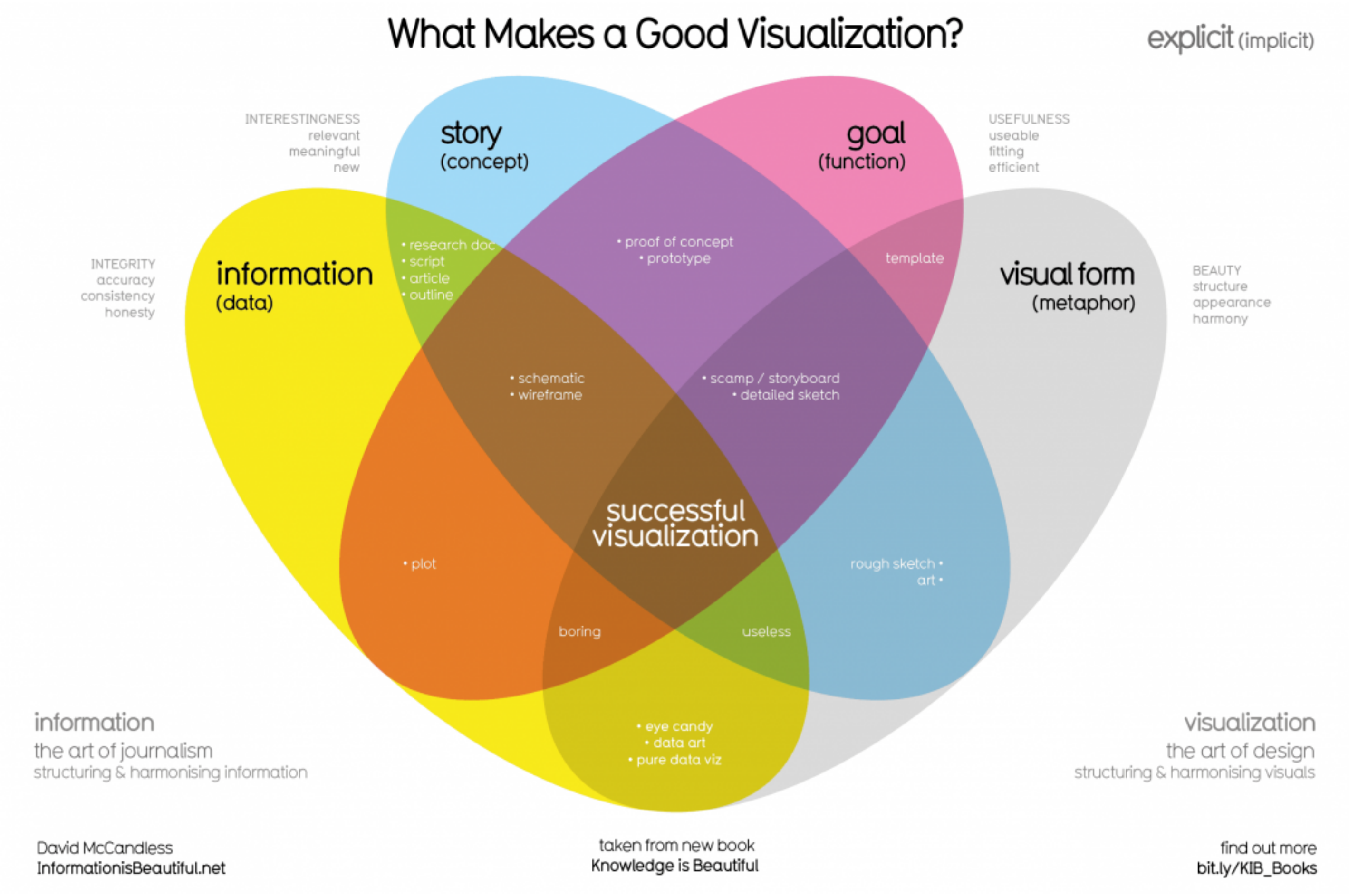
We define migration as a member changing their location on their LinkedIn profile. To develop the list of cities with the most migration, we analyzed migration of LinkedIn members in and out of 50 of the largest U.S. cities (in terms of LI membership) for the past 12 months. So for every 10,000 LinkedIn members in Austin, 548 arrived or departed in the past 12 months.

Good visualization!

This map shows clearly the cities with the most migration. The bubble size and color help the readers see the variances between migration numbers. The note below the map also helps reader understand how the ranking is calculated.

Example2.

Source: https://visualmatters.com/top-17-data-visualizations-review-2017/ (https://visualmatters.com/top-17-data-visualizations-review-2017/)



This graph displays the elements of a successful visualization. The viewers could use gragh to evaluate their own visualizations.

Q4. EDA

EDA is essentially before the analysis and model development. It helps the research process by finding some interesting patterns and questions.

Q5. Tukey1980

Confirmatory: Tukey thinks confirmatory analysis is the analysis and result to confirm or test the idea or hypothesis. One example is running a regression to see whether advertising is a predictive feature for sales.

Exploratory: Tukey thinks exploratoty data analysis is more of an attitude and flexibility, combined with some graph paper. One example is making boxplot of prices across different product categories.