Notes: MS 204 Chapter 1 part III

Overview

- Categorical Data & visualizations
- Bivariate visualizations & linear regression

Categorical data

Ex: X = carrier, Y = origin

```
library(tidyverse); library(oilabs); library(mosaic)
data(nycflights)
nycflights %>% select(carrier, origin) %>% head(3)
```

```
## # A tibble: 3 x 2
##   carrier origin
##     <chr>  <chr>
## 1      VX    JFK
## 2      DL    JFK
## 3      DL    JFK
```

```
tally(~ origin, data = nycflights)
```

```
## origin
##   EWR   JFK   LGA
## 11771 10897 10067
```

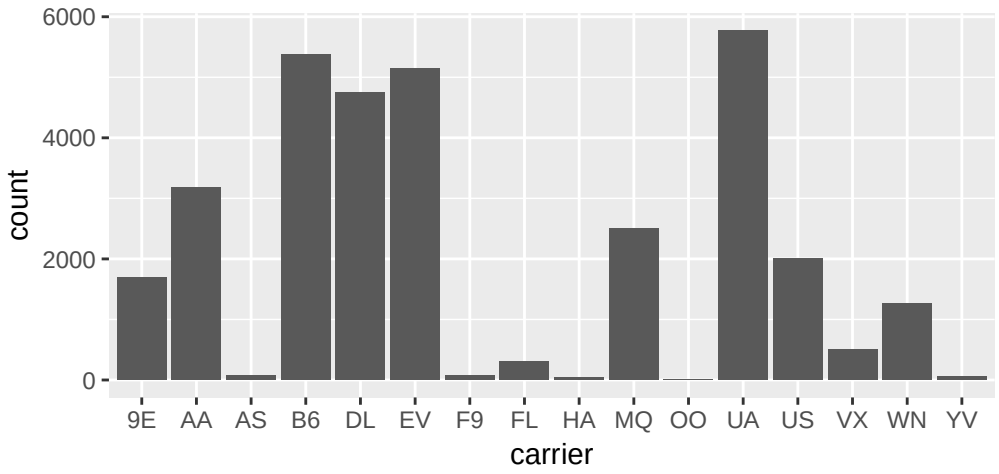
```
tally(origin ~ carrier, data = nycflights, margins = TRUE)
```

```
##        carrier
## origin    9E   AA   AS   B6   DL   EV   F9   FL   HA   MQ   OO   UA   US
##   EWR    121  350   66  625  445 4170    0    0    0  210    1 4559  444
##   JFK   1314 1388    0 4166 2070  118    0    0   34  717    0  440  302
##   LGA    261 1450    0  585 2236  854   69  307    0 1580    2  771 1269
##   Total 1696 3188   66 5376 4751 5142   69  307   34 2507    3 5770 2015
##        carrier
## origin    VX   WN   YV
##   EWR    149  631    0
##   JFK    348    0    0
##   LGA      0  630   53
##   Total  497 1261   53
```

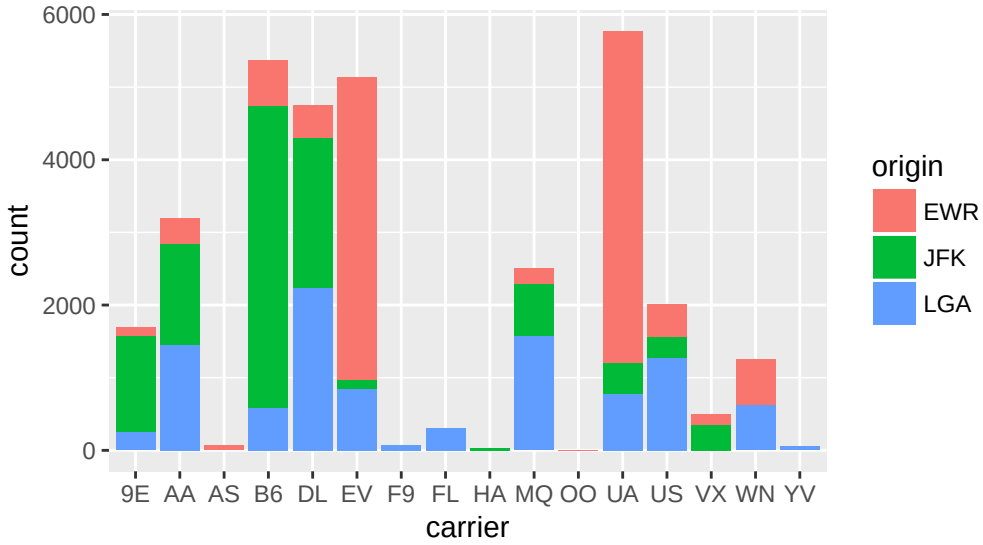
1. P(X = "AA")
2. P(Y = "LGA")
3. P(Y = "LGA" | X = "AA")
4. P(X = "AA" | Y = "LGA")
5. P(X = "AA", Y = "LGA")

Visualizing categorical data

```
qplot(x = carrier, data = nycflights)
```




```
ggplot(aes(x = carrier, fill = origin), data = nycflights) +
  geom_bar()
```



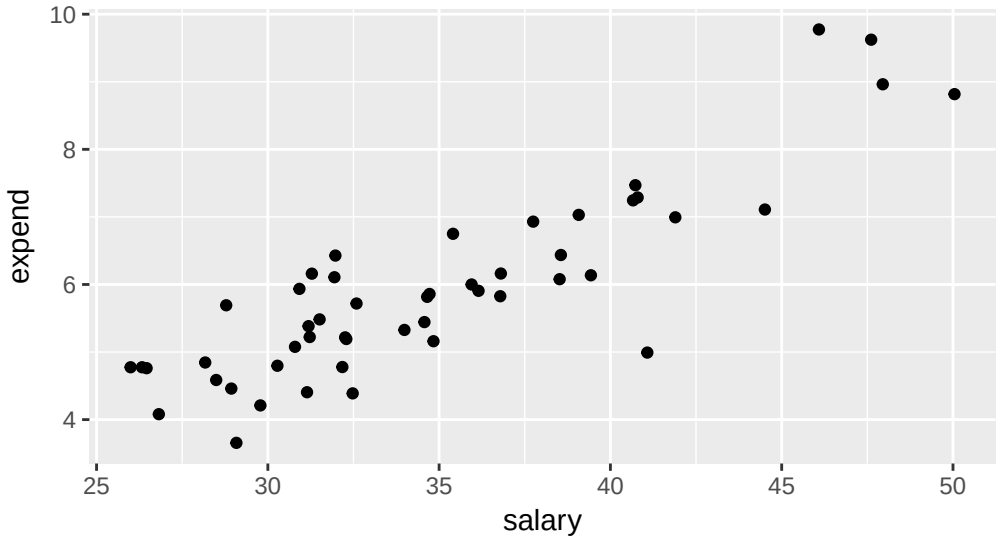
Aside What other visualizations would have been possible with this data set?

Bivariate relationships

```
SAT %>% head()
```

```
##        state expend ratio salary frac verbal math  sat
## 1    Alabama  4.405  17.2 31.144    8    491  538 1029
## 2     Alaska  8.963  17.6 47.951   47    445  489  934
## 3    Arizona  4.778  19.3 32.175   27    448  496  944
## 4   Arkansas  4.459  17.1 28.934    6    482  523 1005
## 5 California  4.992  24.0 41.078   45    417  485  902
## 6   Colorado  5.443  18.4 34.571   29    462  518  980
```

```
qplot(x = salary, y = expend, data = SAT)
```



Correlation

Aside: Identify pairs of variables with a correlation coefficient of -0.9, -0.5, 0, 0.5 and 0.9

Fitting a line

```
SAT %>% head()
```

```
##        state expend ratio salary frac verbal math  sat
## 1    Alabama  4.405  17.2 31.144    8    491  538 1029
## 2     Alaska  8.963  17.6 47.951   47    445  489  934
## 3    Arizona  4.778  19.3 32.175   27    448  496  944
## 4   Arkansas  4.459  17.1 28.934    6    482  523 1005
## 5 California  4.992  24.0 41.078   45    417  485  902
## 6   Colorado  5.443  18.4 34.571   29    462  518  980
```

```
SAT %>% summarize(cor.sat = cor(salary, expend),
                  mean.salary = mean(salary),
                  sd.salary = sd(salary),
                  mean.expend = mean(expend),
                  sd.expend = sd(expend))
```

```
##     cor.sat mean.salary sd.salary mean.expend sd.expend
## 1 0.8698015    34.82892  5.941265     5.90526  1.362807
```

- Interpretations