Lab 05 - Wrangling spatial data

Tensorflow2.0

10/08/2020

Load packages

```
library(tidyverse)
library(dsbox)
```

Exercise 1

First, lets see the number of Denny's in Alaska.

```
dn_ak <- dennys %>%
  filter(state == "AK")
nrow(dn_ak)
```

```
## [1] 3
```

There are 3 Denny's in Alaska.

Exercise 2

Now, lets see the number of La Quinta in Alaska.

```
lq_ak <- laquinta %>%
  filter(state == "AK")
nrow(lq_ak)
```

```
## [1] 2
```

There are 2 La Quinta in Alaska.

Exercise 3

There are 6 pairings between Denny's and La Quinta location in Alaska.

Exercise 4

Now, lets join the two datasets.

```
dn_lq_ak <- full_join(dn_ak, lq_ak, by = "state")
dn_lq_ak
```

```
## # A tibble: 6 x 11
##   address.x city.x state zip.x longitude.x latitude.x address.y city.y zip.y
##   <chr>     <chr>  <chr> <chr>       <dbl>      <dbl> <chr>     <chr>  <chr>
## 1 2900 Den~ Ancho~ AK    99503       -150.       61.2 3501 Min~ "\nAn~ 99503
## 2 2900 Den~ Ancho~ AK    99503       -150.       61.2 4920 Dal~ "\nFa~ 99709
## 3 3850 Deb~ Ancho~ AK    99508       -150.       61.2 3501 Min~ "\nAn~ 99503
## 4 3850 Deb~ Ancho~ AK    99508       -150.       61.2 4920 Dal~ "\nFa~ 99709
## 5 1929 Air~ Fairb~ AK    99701       -148.       64.8 3501 Min~ "\nAn~ 99503
## 6 1929 Air~ Fairb~ AK    99701       -148.       64.8 4920 Dal~ "\nFa~ 99709
## # ... with 2 more variables: longitude.y <dbl>, latitude.y <dbl>
```

There are 6 observations in the joined `dn_lq_ak` dataframe. The name of the variables are address.x, city.x, state, zip.x, longitude.x, latitude.x, address.y, city.y, zip.y, longitude.y, latitude.y.

Exercise 5

We use mutate() functions to add a new variable to data frame while keeping the existing variables.

Now let's define the haversine function in our project.

```
haversine <- function(long1, lat1, long2, lat2, round = 3) {
  # convert to radians
  long1 = long1 * pi / 180
  lat1  = lat1  * pi / 180
  long2 = long2 * pi / 180
  lat2  = lat2  * pi / 180

  R = 6371 # Earth mean radius in km

  a = sin((lat2 - lat1)/2)^2 + cos(lat1) * cos(lat2) * sin((long2 - long1)/2)^2
  d = R * 2 * asin(sqrt(a))

  return( round(d,round) ) # distance in km
}
```

Exercise 6

Now, adding the distance variable in a combined dataset.

```
dn_lq_ak <- dn_lq_ak %>%
  mutate(
    distance = haversine(long1 = longitude.x, lat1 = latitude.x,
                         long2 = longitude.y, lat2 = latitude.y)
  )
dn_lq_ak
```

```
## # A tibble: 6 x 12
##   address.x city.x state zip.x longitude.x latitude.x address.y city.y zip.y
##   <chr>     <chr>  <chr> <chr>       <dbl>      <dbl> <chr>     <chr>  <chr>
## 1 2900 Den~ Ancho~ AK    99503       -150.       61.2 3501 Min~ "\nAn~ 99503
## 2 2900 Den~ Ancho~ AK    99503       -150.       61.2 4920 Dal~ "\nFa~ 99709
## 3 3850 Deb~ Ancho~ AK    99508       -150.       61.2 3501 Min~ "\nAn~ 99503
## 4 3850 Deb~ Ancho~ AK    99508       -150.       61.2 4920 Dal~ "\nFa~ 99709
## 5 1929 Air~ Fairb~ AK    99701       -148.       64.8 3501 Min~ "\nAn~ 99503
## 6 1929 Air~ Fairb~ AK    99701       -148.       64.8 4920 Dal~ "\nFa~ 99709
```

```
## # ... with 3 more variables: longitude.y <dbl>, latitude.y <dbl>,
## #   distance <dbl>
```

Exercise 7

Now, let's find the minimum distance between Denny's and La Quinta for each Denny's location in Alaska

```
dn_lq_ak_mindist <- dn_lq_ak %>%
  group_by(address.x) %>%
  summarise(min_distance = min(distance))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
dn_lq_ak_mindist
```

```
## # A tibble: 3 x 2
##   address.x        min_distance
##   <chr>                   <dbl>
## 1 1929 Airport Way         5.20
## 2 2900 Denali              2.04
## 3 3850 Debarr Road         6.00
```

Exercise 8

Now, let's find the distribution of the distances between Denny's and the nearest La Quinta locations in Alaska

Including the summary statistics.

```
summary(dn_lq_ak_mindist$min_distance)
```

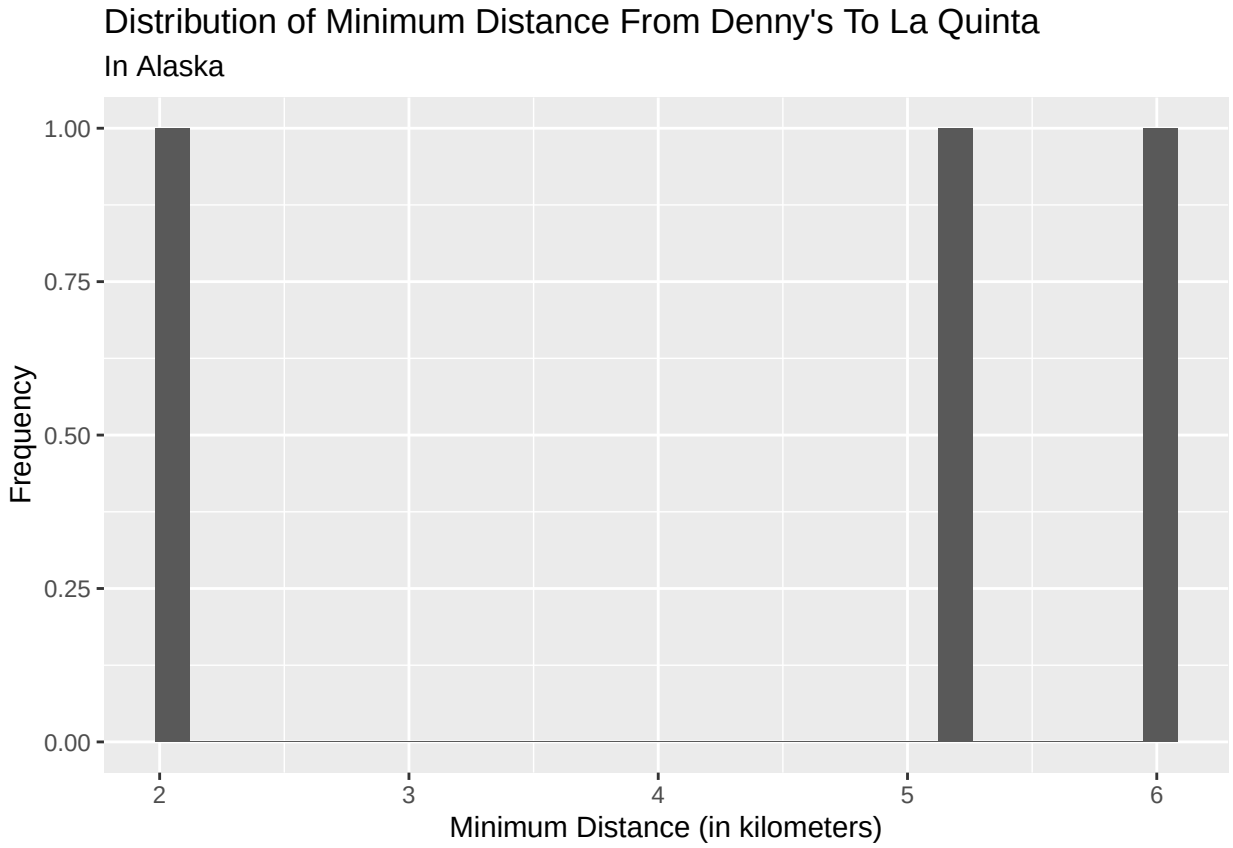
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.035   3.616   5.197   4.410   5.598   5.998
```

It is evident from the summary table above that on average, every nearest La Quinta is separated from Denny's by 4.41 kilometers.

Now, lets visualize the spread of minimum distances.

```
ggplot(data = dn_lq_ak_mindist, aes(x = min_distance)) +
  geom_histogram() +
  labs(
    x = "Minimum Distance (in kilometers)",
    y = "Frequency",
    title = "Distribution of Minimum Distance From Denny's To La Quinta",
    subtitle = "In Alaska"
  )
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The minimum distances are distributed without any single definitive peak. All the Denny's in Alaska have the nearest La Quinta in no more than 6 kilometers, with majority of them having the nearest Laquinta in between 5 and 6 kilometers.

Exercise 9

Repeating the analysis for New Jersey.

```
dn_nj <- dennys %>%
  filter(state == "NJ")

lq_nj <- laquinta %>%
  filter(state == "NJ")

dn_lq_nj <- full_join(dn_nj, lq_nj, by = "state")

dn_lq_nj <- dn_lq_nj %>%
  mutate(
    distance = haversine(long1 = longitude.x, lat1 = latitude.x,
                         long2 = longitude.y, lat2 = latitude.y)
  )

dn_lq_nj_mindist <- dn_lq_nj %>%
  group_by(address.x) %>%
  summarise(min_distance = min(distance))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

Including the summary statistics.

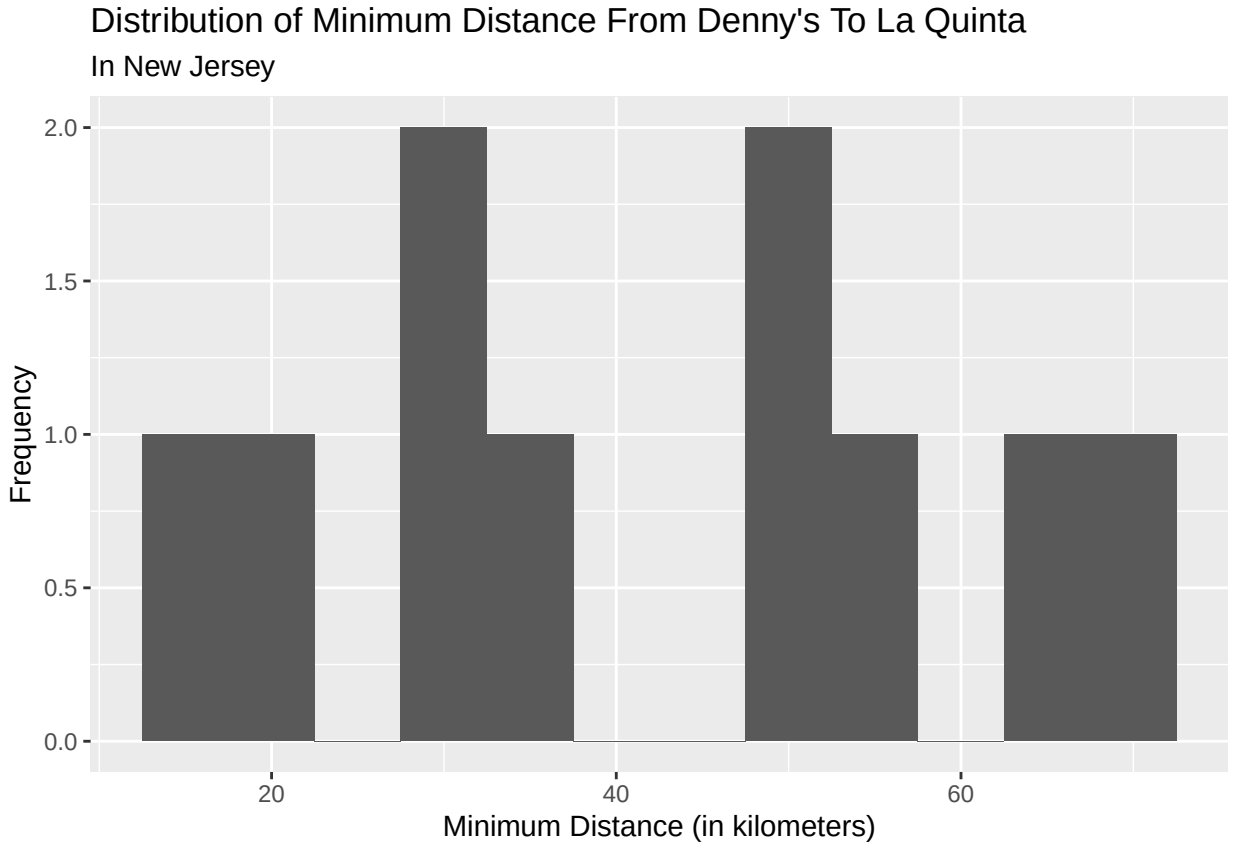
```
summary(dn_lq_nj_mindist$min_distance)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.39   30.88   41.52   42.29   54.40   69.13
```

It is evident from the summary table above that on average, every nearest La Quinta is separated from Denny's by 42.29 kilometers.

Now, lets visualize the spread of minimum distances.

```
ggplot(data = dn_lq_nj_mindist, aes(x = min_distance)) +
  geom_histogram(binwidth = 5) +
  labs(
    x = "Minimum Distance (in kilometers)",
    y = "Frequency",
    title = "Distribution of Minimum Distance From Denny's To La Quinta",
    subtitle = "In New Jersey"
  )
```



The distribution is bimodal, with the minimum distances spread in between 10 and 80 kilometers and two prominent peaks distributed almost symmetrically from 40 on the either side, suggesting that on average minimum distances between Denny's and La Quinta lie somewhere around 40 kilometers.

Exercise 10

Repeating the analysis for Texas.

```r
dn_tx <- dennys %>%
  filter(state == "TX")

lq_tx <- laquinta %>%
  filter(state == "TX")

dn_lq_tx <- full_join(dn_tx, lq_tx, by = "state")

dn_lq_tx <- dn_lq_tx %>%
  mutate(
    distance = haversine(long1 = longitude.x, lat1 = latitude.x,
                         long2 = longitude.y, lat2 = latitude.y)
  )

dn_lq_tx_mindist <- dn_lq_tx %>%
  group_by(address.x) %>%
  summarise(min_distance = min(distance))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

Including the summary statistics.

```r
summary(dn_lq_tx_mindist$min_distance)
```

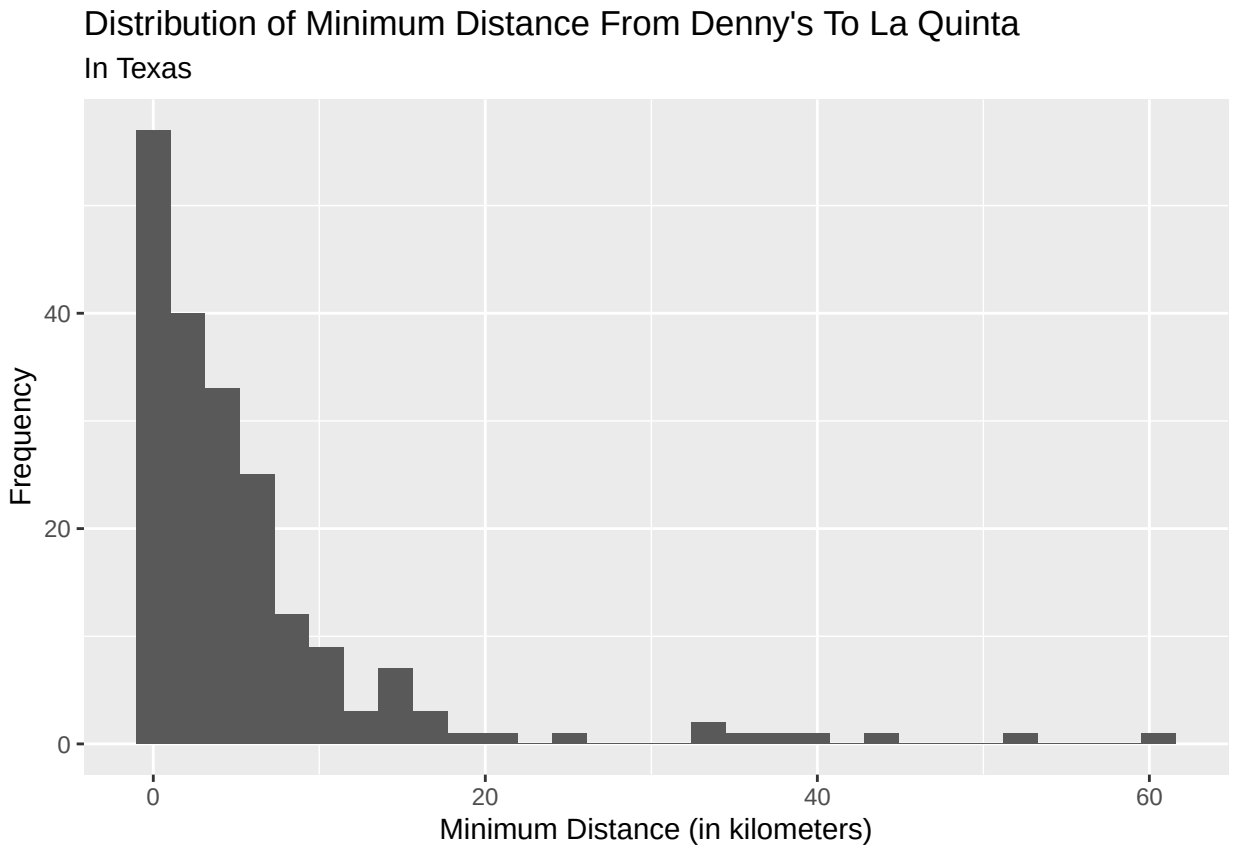
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0160  0.7305  3.3715  5.7918  6.6303 60.5820
```

It is evident from the summary table above that on average, every nearest La Quinta is separated from Denny's by 5.79 kilometers.

Now, lets visualize the spread of minimum distances.

```r
ggplot(data = dn_lq_tx_mindist, aes(x = min_distance)) +
  geom_histogram() +
  labs(
    x = "Minimum Distance (in kilometers)",
    y = "Frequency",
    title = "Distribution of Minimum Distance From Denny's To La Quinta",
    subtitle = "In Texas"
  )
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The distribution is unimodal and right-skewed. This nature of skew supports the data from the summary table above that the median is less than its mean. The distribution also suggests that on average, the minimum distances are clustered in between 0-5 kilometers with only a handful of them being as high as 60 kilometers.

Exercise 11

Repeating the analysis for California.

```
dn_ca <- dennys %>%
  filter(state == "CA")

lq_ca <- laquinta %>%
  filter(state == "CA")

dn_lq_ca <- full_join(dn_ca, lq_ca, by = "state")

dn_lq_ca <- dn_lq_ca %>%
  mutate(
    distance = haversine(long1 = longitude.x, lat1 = latitude.x,
                         long2 = longitude.y, lat2 = latitude.y)
  )

dn_lq_ca_mindist <- dn_lq_ca %>%
  group_by(address.x) %>%
  summarise(min_distance = min(distance))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

Including the summary statistics.

```
summary(dn_lq_ca_mindist$min_distance)
```

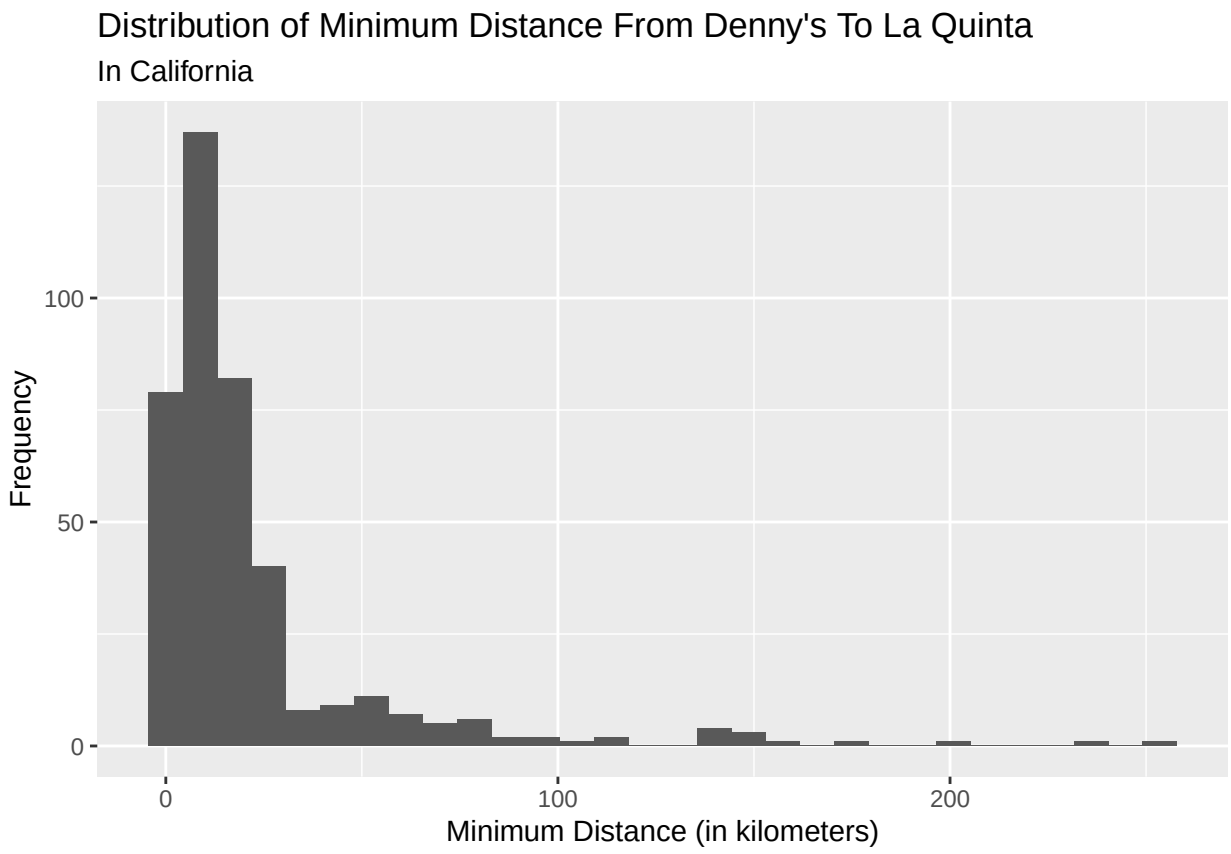
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.016   5.767  11.897  22.083  22.796 253.462
```

It is evident from the summary table above that on average, every nearest La Quinta is separated from Denny's by 22.08 kilometers.

Now, lets visualize the spread of minimum distances.

```
ggplot(data = dn_lq_ca_mindist, aes(x = min_distance)) +
  geom_histogram() +
  labs(
    x = "Minimum Distance (in kilometers)",
    y = "Frequency",
    title = "Distribution of Minimum Distance From Denny's To La Quinta",
    subtitle = "In California"
  )
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The distribution is unimodal with a majority of minimum distances clustered in between 0 and 25 kilometers. However, the minimum distances are far more spread than any other states we have analyzed with some being as high as slightly above 250 kilometers.

Exercise 12

Looking at the mean of the minimum distances from Denny's to La Quinta in different states, it appears that on average the nearest La Quinta are separated to Denny's by the least amount in Alaska. While Texas is also another great candidate to support Hedberg's joke (it has the second least mean and the least median), however due to the presence of outliers (Denny's with minimum distance that is far more than the average), we conclude Alaska as the state where Mitch Hedberg's joke will most likely hold true.