Exploratory Data Analysis (EDA) Paper

Anthony Hu, Rishabh Sud, Aysha Ahmed

06/16/2021

Data Import and Join

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.2
```

```
## -- Attaching packages --------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 3.6.2
```

```
## Warning: package 'tibble' was built under R version 3.6.2
```

```
## Warning: package 'tidyr' was built under R version 3.6.2
```

```
## Warning: package 'readr' was built under R version 3.6.2
```

```
## Warning: package 'purrr' was built under R version 3.6.2
```

```
## Warning: package 'dplyr' was built under R version 3.6.2
```

```
## Warning: package 'forcats' was built under R version 3.6.2
```

```
## -- Conflicts ------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Data

Chicago Census

Key to the variable names of Chicago Census

Chicago socionomic data

Chicago public schools data

```
Chicago_census <- read.csv(url("https://datahub.cmap.illinois.gov/dataset/1d2dd970-f0a6-4736-96a1-3caeb

Chicago_socionomic <- read.csv(url("https://data.cityofchicago.org/api/views/kn9c-c2s2/rows.csv"))

Chicago_schools <- read.csv(url("https://data.cityofchicago.org/api/views/9xs2-f89t/rows.csv"))
```

```
Chicago_combined <- Chicago_socionomic %>%
  rename(GEOG = "COMMUNITY.AREA.NAME") %>%
  left_join(Chicago_census, by = "GEOG")
```

Visualize Categorical-Categorical Relation

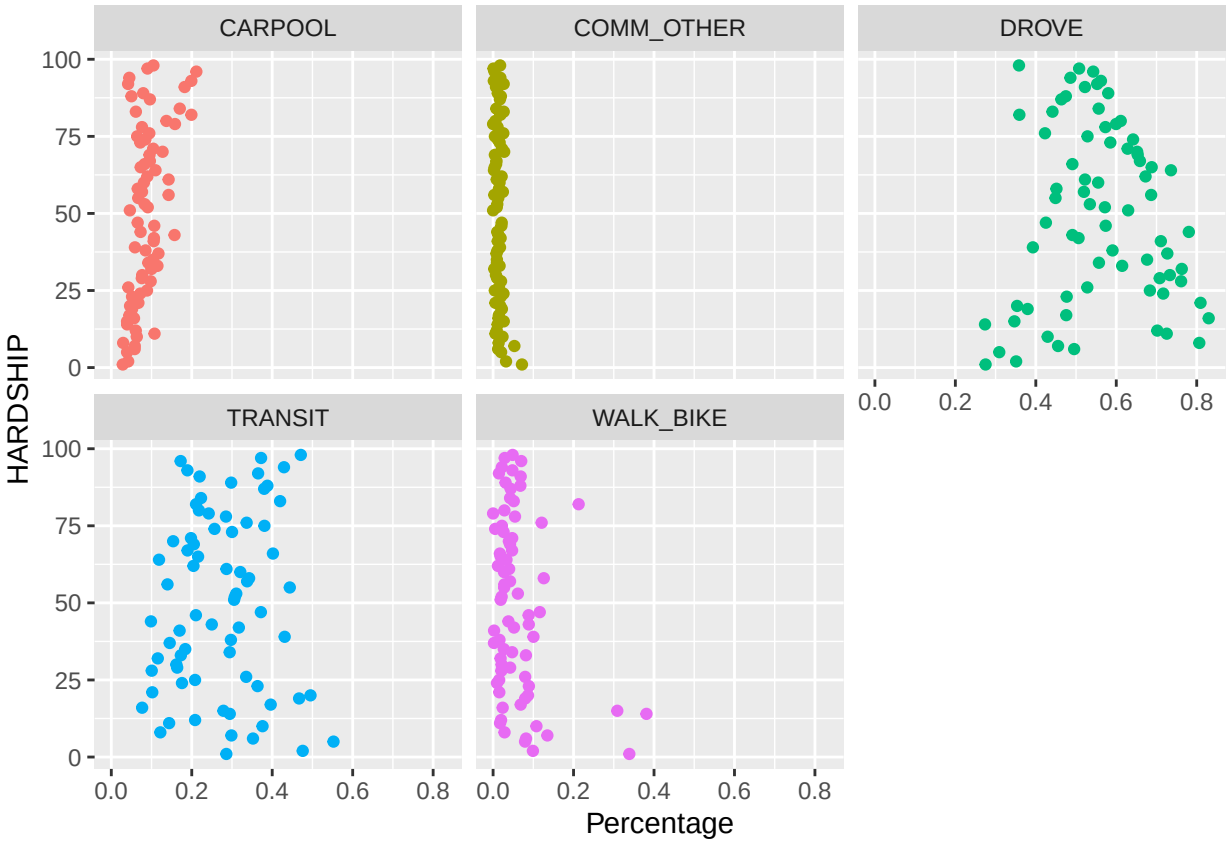
Research Question: What is the relationship between type of transit (using percentage) and the hardship index of each neighborhood.

Graph:

```
Q3 <- Chicago_combined %>%
  select(c("GEOG", "TOT_COMM", "DROVE_AL", "CARPOOL", "TRANSIT",
           "WALK_BIKE", "COMM_OTHER", "HARDSHIP.INDEX" )) %>%
  transmute(GEOG = GEOG, DROVE = DROVE_AL / TOT_COMM, CARPOOL = CARPOOL / TOT_COMM,
            TRANSIT = TRANSIT / TOT_COMM, WALK_BIKE = WALK_BIKE/ TOT_COMM,
            COMM_OTHER = COMM_OTHER / TOT_COMM, HARDSHIP = HARDSHIP.INDEX) %>%
  pivot_longer(c("DROVE", "CARPOOL", "TRANSIT", "WALK_BIKE", "COMM_OTHER"), names_to = "TypeTransit", va

ggplot(data = Q3, mapping = aes(x = Percentage, y = HARDSHIP, color =TypeTransit)) +
  geom_point() +
  facet_wrap(TypeTransit~.) +
  theme(legend.position = "none")
```

```
## Warning: Removed 25 rows containing missing values (geom_point).
```



Description (1-sentence):

In general, there appears to be a strong positive trend between (neighborhood)hardship index and percentage of those who carpool as transit, a horseshoe/quadratic relationship between those who drove to transit, and the hardship index of neighborhoods, an unrelated trend between transit and hardship index, and an unclear relationship between walking and biking, and the hardship index.

Visualize Categorical-Continuous Relation

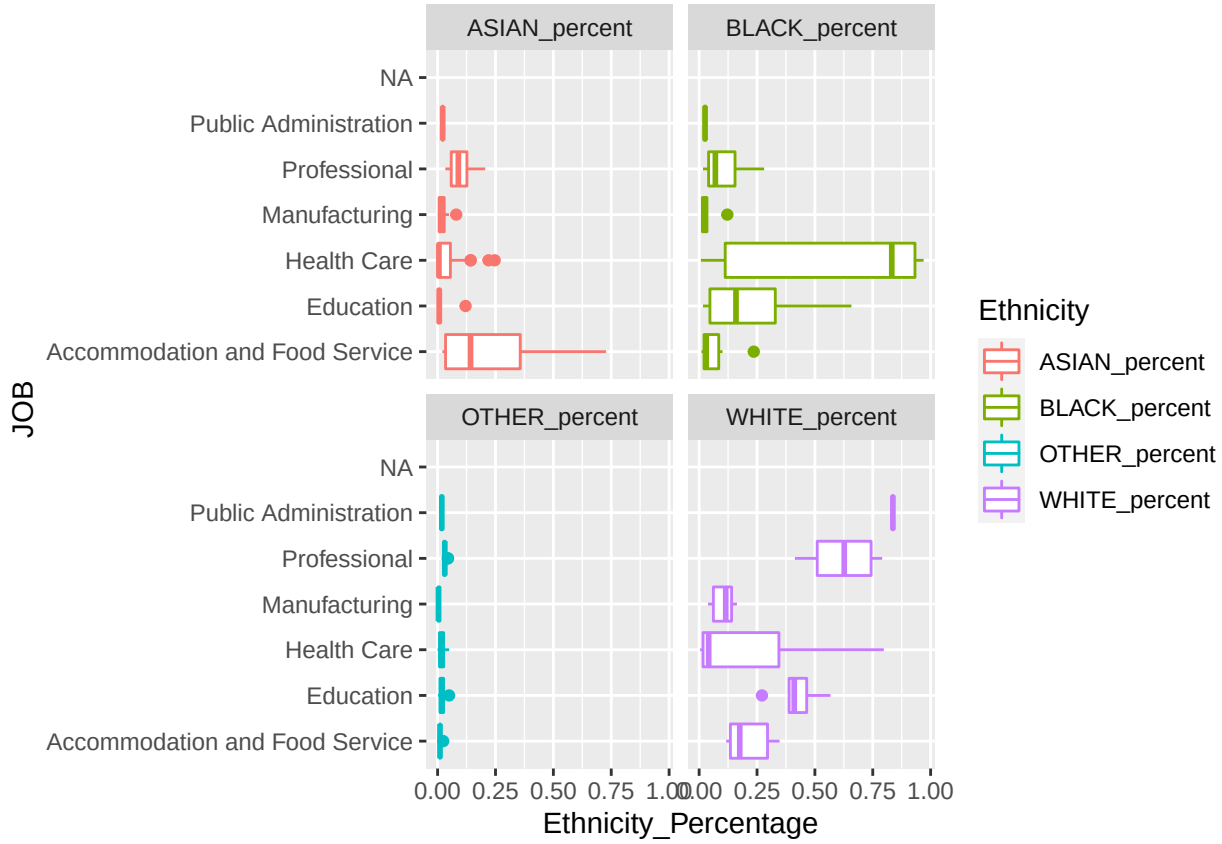
Research Question: What is the relationship between percentage of each ethnicity and most common career industry within each neighborhood.

Graph:

```
Q2_data_alt <-  Chicago_combined %>%
  select(c("TOT_POP", "WHITE", "BLACK", "ASIAN", "OTHER",
           "RES_NAICS1_TYPE")) %>%
  transmute(WHITE_percent = WHITE / TOT_POP, BLACK_percent = BLACK / TOT_POP,
            ASIAN_percent = ASIAN / TOT_POP, OTHER_percent = OTHER/ TOT_POP,
            JOB = RES_NAICS1_TYPE) %>%
  pivot_longer(c("WHITE_percent", "BLACK_percent", "ASIAN_percent", "OTHER_percent"), names_to = "Ethni

ggplot(data=Q2_data_alt) +
  geom_boxplot(aes(x=JOB,y=Ethnicity_Percentage,color=Ethnicity)) +
  facet_wrap(Ethnicity~.) +
  coord_flip()
```

```
## Warning: Removed 20 rows containing non-finite values (stat_boxplot).
```



Description (1-sentence): This breakdown of ethnicity percentage vs occupation shows Asian residents of each community occupying more accommodation and food service roles and Black and White residents occupying more health care roles.

Visualize Continuous-Continuous Relation

Research Question: Within each neighborhood, what is the relationship between the percentage without High School Diplomas and Households below poverty level (grouped by quantiles of income-levels) **Graph:**

```
Q1 <- Chicago_combined %>%
  select(c("PERCENT.AGED.25..WITHOUT.HIGH.SCHOOL.DIPLOMA", "PERCENT.HOUSEHOLDS.BELOW.POVERTY",
           "GEOG", "PER.CAPITA.INCOME")) %>%
  mutate(y = PERCENT.HOUSEHOLDS.BELOW.POVERTY,
         z = PER.CAPITA.INCOME,
         x = PERCENT.AGED.25..WITHOUT.HIGH.SCHOOL.DIPLOMA,
         IncomeRank = ntile(z,5)) %>%
  group_by(IncomeRank)
#Plot

P1 <- ggplot(data = Q1, mapping = aes(x = x,y = y,color = IncomeRank)) +
  geom_point() +
  xlab("Percent aged 25 without a High School Diploma (%)") +
```

```
  ylab("Percent of Households below Poverty (%)") +
  geom_smooth(method = "loess", mapping = aes(x = x,y = y), se = FALSE)
P1
```

```
## `geom_smooth()` using formula 'y ~ x'
```

Description (1-sentence): It shows a highly variable relationship, in which a linear, quadratic, or 3rd degree fits do not best account for the pattern. No clear distinction can be made, but between 0 to 30% of those aged 25 without a Highschool Diploma, there seems to be a weak positive correlation for Percent of Households below Poverty.