Natalie Weger JOUR472 Final Project

Data Analysis Project Part 2

Natalie Weger | JOUR472 | 05/09/2024 | Final Project

```
#loading libraries
library(rvest)
library(tidyverse)
library(janitor)
library(dplyr)
```

1) What are the Baltimore jobs are related directly to the port and what are their median income?

This question is newsworthy because it presents the question of how different types of jobs and incomes will be impacted in light of the Baltimore bridge collapse.

The first dataset that I've used is called the Baltimore Port Jobs, which I took from The 2023 Economic Impact of the Port of Baltimore in Maryland. I used Adobe Acrobat to extract the data from the website and then download it into a CSV file. The data shows the number of total direct jobs that are associated with the Baltimore port. The second dataset that I've collected is the Occupational Data that is provided by the U.S. Bureau of Labor Statistics. This dataset is put into a CSV that will show the median income of different types of jobs in Maryland.

To address the question of how the Baltimore bridge collapse could have potentially affected jobs in the port, I cleaned both datasets and joined them to create a table that shows the amount of direct jobs affected by the port, and the (most likely) median income of these jobs. This step has allowed me to set up my data for this project. However, certain job data within this analysis (which was included in the report, but not in the datasets it provided) included job data with banks, law firms and insurance companies.

Summary Explanation of Variables:

The Baltimore Port Jobs dataset shows the amount of types of jobs that are directly related to the Baltimore Port. The dataset showcases the type of jobs, jobs related directly to the port, Maryland Port Administration jobs related to the port and jobs within the private terminal relate to the port. Here are the variables of the job types and what their interpretation is, according to the report and background research.

- Rail - rail jobs include transportation firms that are involved with railroads, railroad employees
- Truck - truck drivers carrying mass goods
- Terminal - terminal operations, particularly associated with handling autos at both State-owned and private terminals.
- ILA Dockworkers - those who work the docks at the port
- Tug Assist - Barge - escorting, docking and undocking in narrow shipping channels
- Pilots - those who are using aircraft to import goods near the port
- Agents - NA (still finding, could be steamship agents or insurance agents)
- Maritime Services Construction - design, engineering, construction of ship vessels
- Freight Forwarders - helps companies make goods ship to final destination
- Warehouse - warehouse workers who help lifting and organizing of goods
- Government - port jobs related to agencies on this list
- Maryland Port Administration - jobs include plumbers, construction engineers, electrical systems managers
- Dependent Coshippers, signees - help oversee the transportation of goods in the port

The Maryland Jobs dataset shows a general overview of Maryland Jobs and shows their state (Maryland), a code for the occupation, occupation title, their hourly and annual mean salary.

Importing and cleaning the data:

```
#import baltimore jobs related to the port, rename the jobs column and then take out unnec
baltimore_port_jobs <- read_csv("baltimore_port_jobs.csv") %>%
clean_names() %>%

#take out unnecessary rows/columns
rename(job_types = 1) %>%
subset(select = -c(5))
baltimore_port_jobs <- slice(baltimore_port_jobs, -c(1,4,16,17))
```

```
#clean the names under the job_titles column
baltimore_port_jobs$job_types <- tolower(gsub("[ /]", "_", baltimore_port_jobs$job_types))

#import the mean incomes of different jobs in maryland
maryland_jobs <- read_csv("maryland_jobs.csv") %>%

#clean the dataset names
clean_names() %>%
rename(state = 1) %>%

#take out columns so that it just displays annual mean income
subset(select = -c(2,4,5,7,8))

#filter out all rows where there is a null value for annual mean data
sum(is.na(maryland_jobs$a_mean))
```

```
[1] 0
```

```
maryland_jobs <- maryland_jobs %>%
  filter(a_mean != "#" & a_mean != "*")
```

Mutating The Data

```
#use mutate to create a new column called occ_title_2 and change the job names from maryla
maryland_jobs <- maryland_jobs %>%
  mutate(
    occ_title_2 = case_when(
      occ_title == "Railroad Brake, Signal, and Switch Operatorand Locomotive Firers"   ~
      occ_title == "Heavy and Tractor-Trailer Truck Drivers"   ~ "truck",
      occ_title == "Captains, Mates, and Pilotof Water Vessels"   ~ "pilots",
      occ_title == "Construction Laborers"   ~ "maritime_services_construction",
      occ_title == "Helpers, Construction Trades, All Other"   ~ "warehouse",
      occ_title == "Shipping, Receiving, and Inventory Clerks"   ~ "dependent_shippers_con
      occ_title == "Laborerand Freight, Stock, and Material Movers, Hand"   ~ "freight_for
      occ_title == "Operating Engineerand Other Construction Equipment Operators" ~ "maryl
      occ_title == "Electrical and ElectronicInstallerand Repairers, Transportation Equipm
      occ_title == "Plumbers, Pipefitters, and Steamfitters" ~ "maryland_port_administrati
      occ_title == "First-Line Supervisorof Construction Tradeand Extraction Workers" ~ "m
      TRUE                 ~ occ_title
    )
  )
```

```
#i couldn't find a little less than half of the names of baltimore port jobs and properly

#now i will calculate the median of maryland port administration jobs by creating a new da
mpa_jobs <- maryland_jobs %>%
  filter(occ_title_2 == "maryland_port_administration")

#filter out commas from the values
  mpa_jobs$a_mean <- as.numeric(gsub(",", "", mpa_jobs$a_mean))

#calculate the median income for the maryland port administration
  maryland_port_median_income <- median(mpa_jobs$a_mean)
  head(maryland_port_median_income)
```

```
[1] 63400
```

```
#the median income of jobs in the maryland port administration is $63,400
```

Joining the data to create a new dataset

```
#creating a dataset from the maryland jobs database with just these new occupations
updated_maryland_jobs <- maryland_jobs %>%
   select(occ_title_2, a_mean) %>%
   filter(occ_title_2 %in% c("rail","truck","pilots","maritime_services_construction", "wa

#create a new dataset with the updated baltimore port jobs that can match the maryland job
updated_baltimore_jobs <- baltimore_port_jobs %>%
   select(job_types, total_direct_jobs) %>%
   filter(job_types %in% c("rail","truck","pilots","maritime_services_construction", "ware

#inner join the updated maryland wages and job amounts, and take out commas in the values
inner_join_table <- updated_maryland_jobs %>%
  inner_join(updated_baltimore_jobs, by = c("occ_title_2" = "job_types"))
  inner_join_table$a_mean <- as.numeric(gsub(",", "", inner_join_table$a_mean))
  head(inner_join_table)
```

```
# A tibble: 6 x 3
  occ_title_2                    a_mean total_direct_jobs
  <chr>                           <dbl>             <dbl>
1 freight_forwarders              35370               673
2 truck                           51090              4794
```

```
3 maritime_services_construction  38980              977
4 dependent_shippers_consignees   38940             3474
5 warehouse                       37370             1237
6 pilots                         102240              139
```

```
#calculate the median income of all maryland jobs?
maryland_median_income <- median(maryland_jobs$a_mean)
#the median income of all maryland jobs is $47,360

#calculating the median income of the baltimore port jobs from combining the inner join ta
updated_baltimore_median_income <- median(c(inner_join_table$a_mean, maryland_port_median_
head(updated_baltimore_median_income)
```

```
[1] 42525
```

```
#the median of jobs related to the baltimore port is $42,525
```

Scraping a website from Glassdoor to get more private salary data about the Baltimore Port

```
#identify the website that we are scraping
sboe_url <- "https://www.glassdoor.com/Salary/Ports-America-Baltimore-Salaries-EI_IE147702

# read in the html
results <- sboe_url %>%
  read_html() %>%
  html_table()
  results
```

```
[[1]]
# A tibble: 15 x 3
   `Job Title`                                 Total PayBase | Addi~1 `Open Jobs`
   <chr>                                       <chr>                  <lgl>
 1 Superintendent11 Salaries submitted$92K-$~ $92K-$131K$103K | $6K   NA
 2 Marine Superintendent3 Salaries submitted~ $105K-$146K$114K | $9K NA
 3 Claims Administrator2 Salaries submitted$~ $43K-$60K$51K | $0     NA
 4 Billing Specialist2 Salaries submitted$44~ $44K-$58K$51K | $0     NA
 5 Claims Adjuster1 Salaries submitted$52K-$~ $52K-$78K$63K | $0     NA
 6 Long Term Substitute Teacher1 Salaries su~ $50K-$79K$63K | $0     NA
 7 Operations Manager1 Salaries submitted$84~ $84K-$125K$93K | $9K   NA
 8 Logistic Coordinator1 Salaries submitted$~ $43K-$57K$49K | $0     NA
 9 Marine Superintendnet1 Salaries submitted~ $104K-$143K$114K | $8K NA
10 Development Leader Program1 Salaries subm~ $99K-$177K$118K | $14K NA
11 Long Shoreman1 Salaries submitted$45K-$74~ $45K-$74K$58K | $0     NA
12 Crane Mechanic1 Salaries submitted$62K-$8~ $62K-$84K$72K | $0     NA
13 Crane Superintendent1 Salaries submitted$~ $98K-$138K$108K | $8K  NA
14 Terminal Operations Coordinator1 Salaries~ $56K-$77K$62K | $3K    NA
15 Executive Assistant1 Salaries submitted$5~ $50K-$71K$60K | $0     NA
# i abbreviated name: 1: `Total PayBase | Additional`
```

```
#making the scrapped website data into a dataset
port_america_data <- data.frame(Job_Title = c("Superintendent11 Salaries submitted$92K-$13
                                  "Marine Superintendent3 Salaries submitted$105K-$146K$11
                                  "Claims Administrator2 Salaries submitted$43K-$60K$51K |
                                  "Billing Specialist2 Salaries submitted$44K-$58K$51K | $
                                  "Claims Adjuster1 Salaries submitted$52K-$78K$63K | $00
                                  "Long Term Substitute Teacher1 Salaries submitted$50K-$7
                                  "Operations Manager1 Salaries submitted$84K-$125K$93K |
                                  "Logistic Coordinator1 Salaries submitted$43K-$57K$49K |
                                  "Marine Superintendnet1 Salaries submitted$104K-$143K$11
                                  "Development Leader Program1 Salaries submitted$99K-$177
                   stringsAsFactors = FALSE)


# Parse the data into separate columns
port_america_data <- port_america_data %>%
  mutate(
    # Extract job title
    Job_Title = sub("\\d+ Salaries submitted.*", "", Job_Title),
    # Extract salary range
    Salary_Range = sub(".*?(\\$\\d+K-\\$\\d+K).*", "\\1", Job_Title),
    # Extract low salary
    Low_Salary = as.numeric(gsub("\\D", "", sub("\\$(\\d+)K-\\$\\d+K.*", "\\1", Salary_Ran
    # Extract high salary
    High_Salary = as.numeric(gsub("\\D", "", sub("\\$\\d+K-\\$(\\d+)K.*", "\\1", Salary_Ra
  ) %>%
  select(Job_Title, Salary_Range, Low_Salary, High_Salary)

# View the parsed data
view(port_america_data)
```

My Question One Findings:

There are about 19,970 direct jobs that are related to the Baltimore Port split into 12 different categories. Of those categories, I was able to match eight of them with median incomes through the Maryland jobs database. I was also able to find jobs related to the Maryland Port Administration through this website, and was able to calculate the median of those jobs to create this finding. However, I was unable to match the other four (terminal, ILA/dockworkers, tug assist/barge, government).

Of the eight Baltimore port jobs I found, the median annual income for those jobs is about $42,525. However, the median annual income of all Maryland jobs is about $47,360. Although the median income of Baltimore port jobs is lower than the average Maryland job median income, this is probably because higher-paying port jobs, such as government and jobs such as banks, law firms and insurance companies were excluded from this calculation. Although I was able to find the exact government agencies that are related to the Baltimore Port, I was still unable to find the exact jobs associated with these agencies and the port. Therefore, I left this calculation out of my project because I didn't want to misrepresent these agencies. This is a great first step in my reporting, and my next step on this pitch would be to email/call these government agencies and clarify the specific jobs that are related to the port. This would likely drive the Baltimore Port median income data much higher.

These results raise the question of how the Baltimore bridge collapse will affect these jobs. Will the annual median income of $42,525 decrease now that a major part of the bridge is gone? These findings could help support a story about this in the future as we get more data on the bridge.

2) What are the direct jobs associated with different commodities related to the port?

This will be newsworthy as it examines the amount of jobs, associated with different commodities, and how it could be affected by the Baltimore bridge collapse. I was also able to find this dataset from The 2023 Economic Impact of the Port of Baltimore in Maryland. I used Adobe Acrobat to extract the data from the website and then download it into a CSV file. I then cleaned the data and arranged it in descending order to show the total direct jobs, by commodity, that are affected by the port.

```
#load baltimore commodity data
bmore_commodities <- read_csv("bmore_commodities.csv") %>%

#clean the data to make the column names lowercase
clean_names() %>%

#remove other columns to just show total direct jobs
subset(select = -c(2,3))
```

```
Rows: 14 Columns: 4
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (1): COMMODITIES
num (3): MPA
DIRECT JOBS, PRIVATE
DIRECT JOBS, TOTAL
DIRECT JOBS

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#take away the ".000" decimals
bmore_commodities$total_direct_jobs <- round(bmore_commodities$total_direct_jobs)

#remove totals row
bmore_commodities <- slice(bmore_commodities, -c(14))

#make commodity values lowercase
bmore_commodities$commodities <- tolower(gsub("[ /]", "_", bmore_commodities$commodities))

#arrange the amount of direct jobs in descending order
bmore_commodities <- bmore_commodities %>%
 arrange(desc(total_direct_jobs))

head(bmore_commodities)
```

```
# A tibble: 6 x 2
  commodities    total_direct_jobs
  <chr>                      <dbl>
1 containers                  7266
2 not_allocated               3478
3 other_dry_bulk              2426
4 automobiles                 2065
5 roro                        1132
6 break_bulk                   794
```

Question Two Findings:

After cleaning my dataset of about 13 commodities with direct jobs that are related to the Baltimore Port, I found that containers are the biggest commodity that has direct jobs related to the port. This commodity has 7266 direct jobs related to the port. Containers, according to the Maryland Port Administration report, are classified as containerized cargo that is passed through the port. This commodity tends to generate the greatest direct job impact with firms in the maritime service sector. Jobs that are impacted by containerized cargo include longshoremen, freight forwarders/customhouse brokers, warehouses, steamship agents, trucking firms and railroads. For the next step of this project, it will be interesting to see how the median incomes of these jobs relate to the container commodity.

Summary Explanation of Variables:

This Baltimore Commodities dataset shows the commodities related to the port and the direct jobs associated with them. The data variables includes the names of these commodities and the number of the total direct jobs associated with them.

3) How many jobs in different Maryland counties are directly affected by the port?

To answer this question, I used data from The 2023 Economic Impact of the Port of Baltimore in Maryland. Again, I used Adobe Acrobat to extract the data from the website and then download it into a CSV file. After cleaning up unnecessary rows in the data, I was able to find amount of direct jobs from the port that are in different counties of Maryland. This will allow me to see the scope of the port's impact and how it affects employment and jobs beyond Baltimore.

Summary Explanation of Variables:

This Port Counties dataset shows the counties that have jobs related to the port. It includes variable data including county names, the share (percentage) of jobs related to the port in each county and the exact number of total direct jobs in each county related to the port.

```
#import county data
port_counties <- read_csv("port_counties.csv") %>%
clean_names()
```

```
Rows: 9 Columns: 3
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (2): JURISDICTION, SHARE
num (1): TOTAL DIRECT JOBS

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
#delete the total row and county row
port_counties <- slice(port_counties, -c(2,9))

#make counties lowercase
port_counties$jurisdiction <- tolower(gsub(" ", "_", port_counties$jurisdiction))

head(port_counties)
```

```
# A tibble: 6 x 3
  jurisdiction   share  total_direct_jobs
  <chr>          <chr>              <dbl>
1 baltimore_city 27.69%              5529
2 anne_arundel   21.46%              4286
3 baltimore      32.79%              6548
4 harford        5.84%               1167
5 howard         1.95%                390
6 other_maryland 7.48%               1493
```

Question Three Findings:

My findings showed that the biggest counties with jobs directly related to the Baltimore Port include Baltimore County with about 32.79% of jobs and then Baltimore City with about 27.69% of jobs. This finding raises the question of how these counties and surrounding countries (such as Anne Arundel and Harford County) will be affected by the Baltimore bridge collapse. I included a data visualization of my findings below.