Homework One Analysis

Link to Github

https://github.com/safiaread/homework-1

Enrollment Data

1. How many observations exist in your current dataset?

```
[1] 13276162
```

2. How many different plan_types exist in the data?

```
# A tibble: 65 × 3
# Groups:   plan_type, year [65]
   plan_type                              year     n
   <chr>                                 <int> <int>
 1 1876 Cost                              2010  1099
 2 1876 Cost                              2011  1377
 3 1876 Cost                              2012  1591
 4 1876 Cost                              2013  1774
 5 1876 Cost                              2014  1808
 6 1876 Cost                              2015  2006
 7 Continuing Care Retirement Community   2010    42
 8 ESRD I                                 2010    14
 9 ESRD II                                2010     2
10 Employer/Union Only Direct Contract PDP 2010  1178
# i 55 more rows
```

3. Provide a table of the count of plans under each plan type in each year. Your table should look something like Table 1.

```
# A tibble: 11 × 7
# Groups:   plan_type [11]
   plan_type    Count_2010 Count_2011 Count_2012 Count_2013 Count_2014 Count_2015
   <chr>             <int>      <int>      <int>      <int>      <int>      <int>
 1 1876 Cost          6035       6851       7633       7731       7069       7157
 2 Employer/U…       28700      28697      28669      25526      25528      25630
 3 HCPP - 183…        3604         11         11         10          9          9
 4 HMO/HMOPOS       506802     528473     507272     530909     523304     479275
 5 Local PPO        417551     515700     636701     633884     664716     704993
 6 MSA                 135       6421       6416       6431       6449       6518
 7 Medicare P…      893609     771694     815223     826907    1122209     991457
```

```
 8 National P…        717       781       858       953      1118      1216
 9 PFFS            385733     45781     36423     31919     24905     13658
10 Pilot              53         3         3         2         2         2
11 Regional P…     24442     22773     21602     19970     19773     17578
```

4.Remove all special needs plans (SNP), employer group plans (eghp), and all "800-series" plans. Provide an updated version of Table 1 after making these exclusions.

```
# A tibble: 8 × 7
# Groups:   plan_type [8]
  plan_type                      `2010` `2011` `2012` `2013` `2014` `2015`
  <chr>                          <int>  <int>  <int>  <int>  <int>  <int>
1 1876 Cost                       4923   5829   6647   6759   6207   6329
2 HMO/HMOPOS                     34460  33931  37551  37179  38893  36588
3 Local PPO                      11652  13874  17030  17089  17169  16728
4 MSA                               68    131    132    145    163    232
5 Medicare Prescription Drug Plan 391205 295458 289044 278091 301082 269153
6 National PACE                    717    781    858    953   1118   1216
7 PFFS                           54119  22038  17449  12945   6053   4232
8 Regional PPO                   10659  10995  11279   9660  10420   8531
```

5. Merge the contract service area data to the enrollment data, and restrict the data only to contracts that are approved in their respective counties. The R script to create the service area dataset is here: Contract Service Area. And you can follow the _BuildFinalData.R script to see where/how I join the datasets. Limiting your dataset only to plans with non-missing enrollment data, provide a graph showing the average number of Medicare Advantage enrollees per county from 2010 to 2015. Be sure to format your graph in a meaningful way.
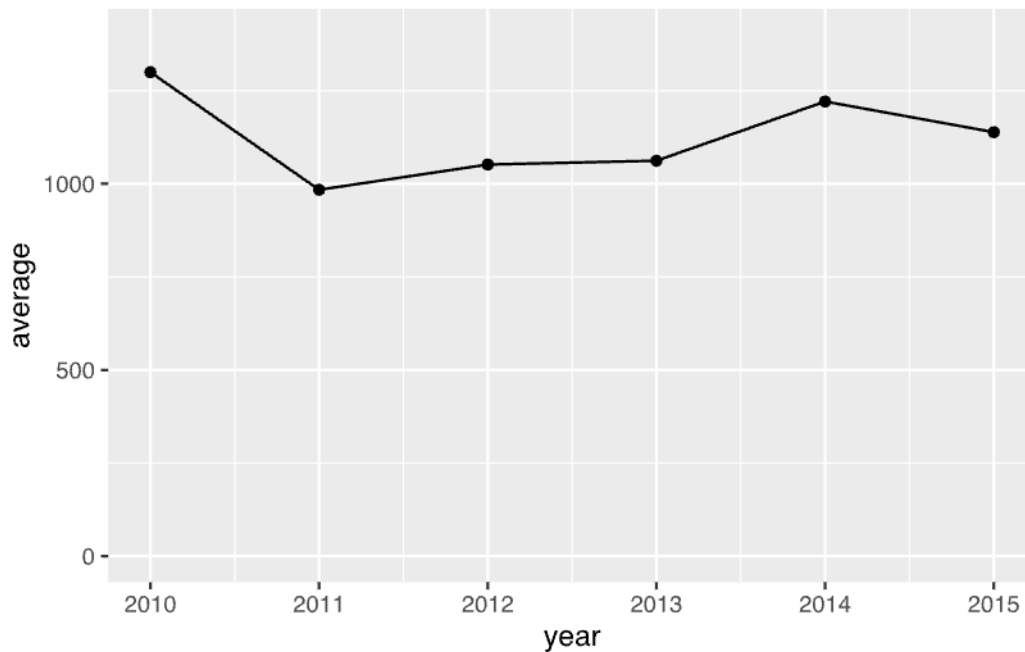
Premium Data

6. Merge the plan characteristics data to the dataset you created in Step 5 above. Note that you'll need to join the Market Penetration Data in order to get the information you need to merge the plan characteristics. This is because the plan characteristics data only have state name and county (not FIPS codes). The penetration files have both FIPS codes and state/county names, so that dataset serves as a good crosswalk file. Provide a graph showing the average premium over time. Don't forget about formatting!

I am still working on how to merge this data. I am specifically having trouble with the premium dataset. #```{r} premiums <- readRDS("data/output/plan_premiums.rds") penetration <- readRDS("data/output/ma_penetration.rds") head(penetration) head(premium)

s_joined <- joined%>% subset(select = c(contractid, fips, ssa, year))

new_joined <- s_joined%>% left_join(penetration, by = c("fips", "ssa", "year"))%>% left_join(premiums, by = c("contractid", "county", "state", "year"))

head(new_joined)

new_joined$premium #%>% #left_join(premium, by = c("contractid", "county", "state"))

avg_premium <- new_joined%>% subset(premium != "NA")%>% group_by(year)%>% summarise(average_premium = mean(premium))

ggplot(avg_premium, aes(x = year, y=average_premium))+ geom_point()+ geom_line()

```
7.
Provide a graph showing the percentage of $0 premium plans over time. Also…
remember to format things.

I need the merged dataset from above to run this code but this is what I assume
I need to do once it is merged.

#```{r}
zero <- new_joined%>%
subset(premium == 0)%>%
group_by(year)%>%
count(zero_n)

all <- ew_joined%>%
group_by(year)%>%
count(all_n)

all%>%
inner_join(zero, by = "year")%>%
mutate(percent = zero_n/all_n)

ggplot(aes(x= year, y = percent))+
geom_point()+
geom_line()

# can use kable or stargazer to make tables look nice.
```

Summary Questions

8. Why did we drop the "800-series" plans?

800 series plans are offered directly to employers and unions.

9. Why do so many plans charge a $0 premium? What does that really mean to a beneficiary?

Maybe they charged $0 because they had really high deductibles. They also could have contracts
with a small network of providers with high out of pocket costs.

10. Briefly describe your experience working with these data (just a few sentences). Tell me one
    thing you learned and one thing that really aggravated you.

With this data, it was definately an adjustment to manipulate it in VSCode. I have previously used
RStudio, and I think you were right that this is a less intuitive interface. I had trouble figuring
out how to view the data to get a grasp of how to manipulate it. I also have never worked with a
dataset so large before, so it took some time to figure out how to upload it to my computer and
organize everything so my computer would run. I'm having some trouble with the merging and
the formating at the momment.