CS595: Big data project

Professor: Dr. Chuck Cartledge.

Presented by: Srinivas Havanur, Kevin Garner, Prasanna Sajjan

Old Dominion University

Individual Responsibilities

- Srinivas Havanur – Analysis, Data Extraction, Menu driven script using bash script, Presentation Slides.
- Kevin Garner – Analysis, Data Extraction, Project documentation, Presentation Slides
- Prasanna Sajjan – Analysis, Data Extraction, Finding correlation value, p value, Generation of scatter plot and venn diagram, Presentation slides

What is the problem and why is it important?

- What is the correlation between the average Medicare billings for the "Cardiovascular stress test" procedure by address and the total cost of pharmaceutical payments made to each address?
- Some pharmaceutical companies may pay physicians to use their products with some procedures that have the highest profit margin.

Technologies used

- Data Extraction was handled using Pig Latin script
- Interactive Menu Driven bash script is written that will call pig script to generate the outputs.
- Correlation value, p value, Scatterplots, Venn diagrams are generated using ‘R GUI’.

Challenges in joining two databases

- Abbreviations

  Eg: 21 YOST BLVD,FOREST HILLS PLZ, STE 216,PITTSBURGH,PA

  21 YOST BOULEVARD,FOREST HILLS PLZ, STE 216,PITTSBURGH,PA

  List of abbreviations handled

  STE, ST, RD, APT, AVE, BLDG, DEPT, LN, PLZ, RDG, DR, PKWY, VLY, PL

- Hashtags

  Eg:

  21 YOST BLVD,FOREST HILLS PLZ, STE #216,PITTSBURGH,PA

  21 YOST BLVD,FOREST HILLS PLZ, STE 216,PITTSBURGH,PA

- Whitespace
- Trailing space
- Addresses with missing information

  Eg: 21 YOST BLVD,FOREST HILLS PLZ, STE #216, PITTSBURGH,PA

  21 YOST BLVD,FOREST HILLS PLZ, 216, PITTSBURGH,PA

Running the script

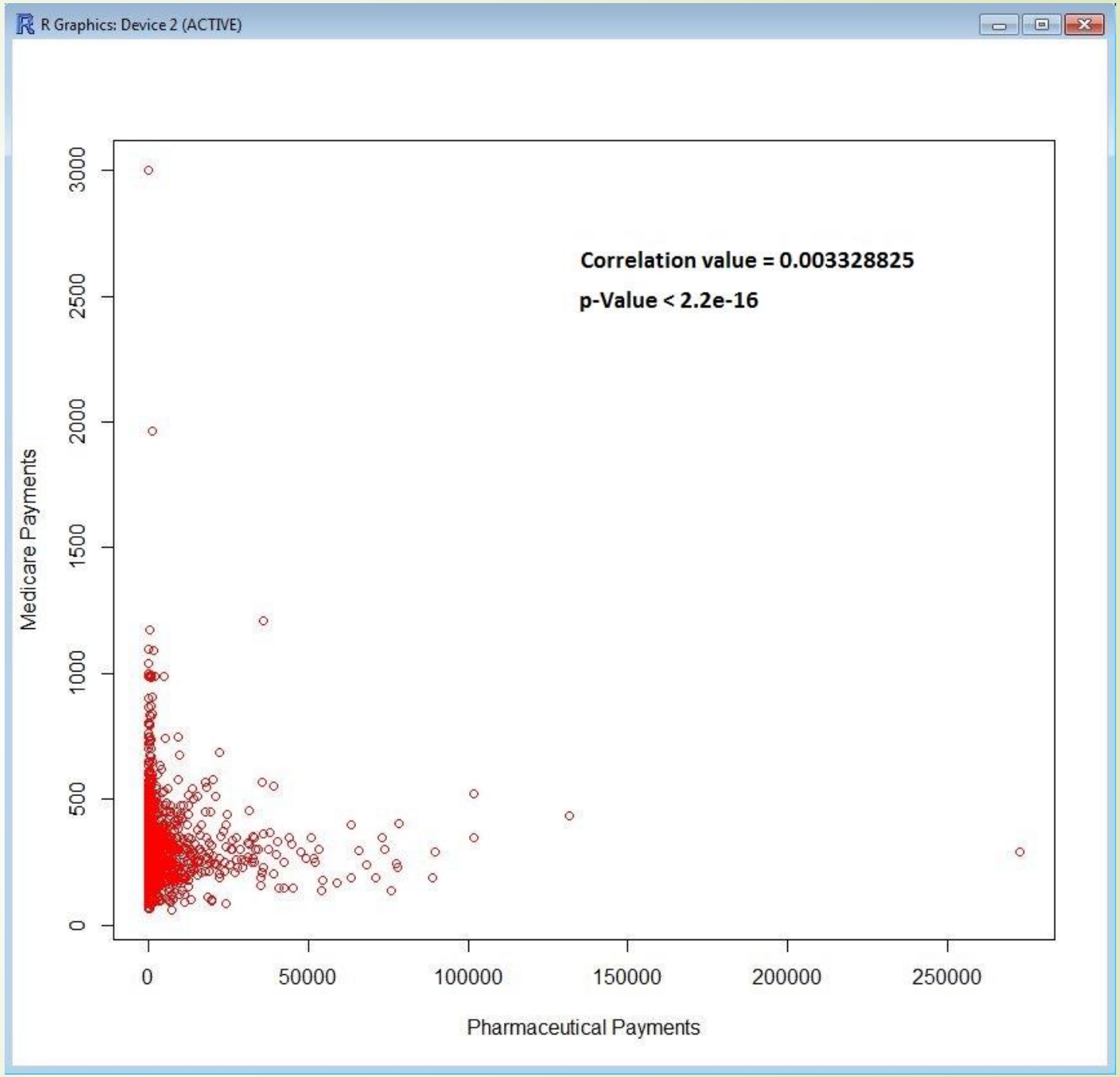
```
##################################################################################################

                                                Project : Bigdata
                                                Dr Chuck Cartledge
                                  By: Srinivas Havanur, Kevin Garner, Prasanna Sajjan
##################################################################################################
1. Cardiovascular stress test(93015)
2. Generate count of medicare and pharmaceutical records before and after the join(93015).
3. Extra Credit. Electrocardiogram report(93010)
4. Generate count of medicare and pharmaceutical records before and after the join(93010).
Please enter your choice
```

Scatter plot

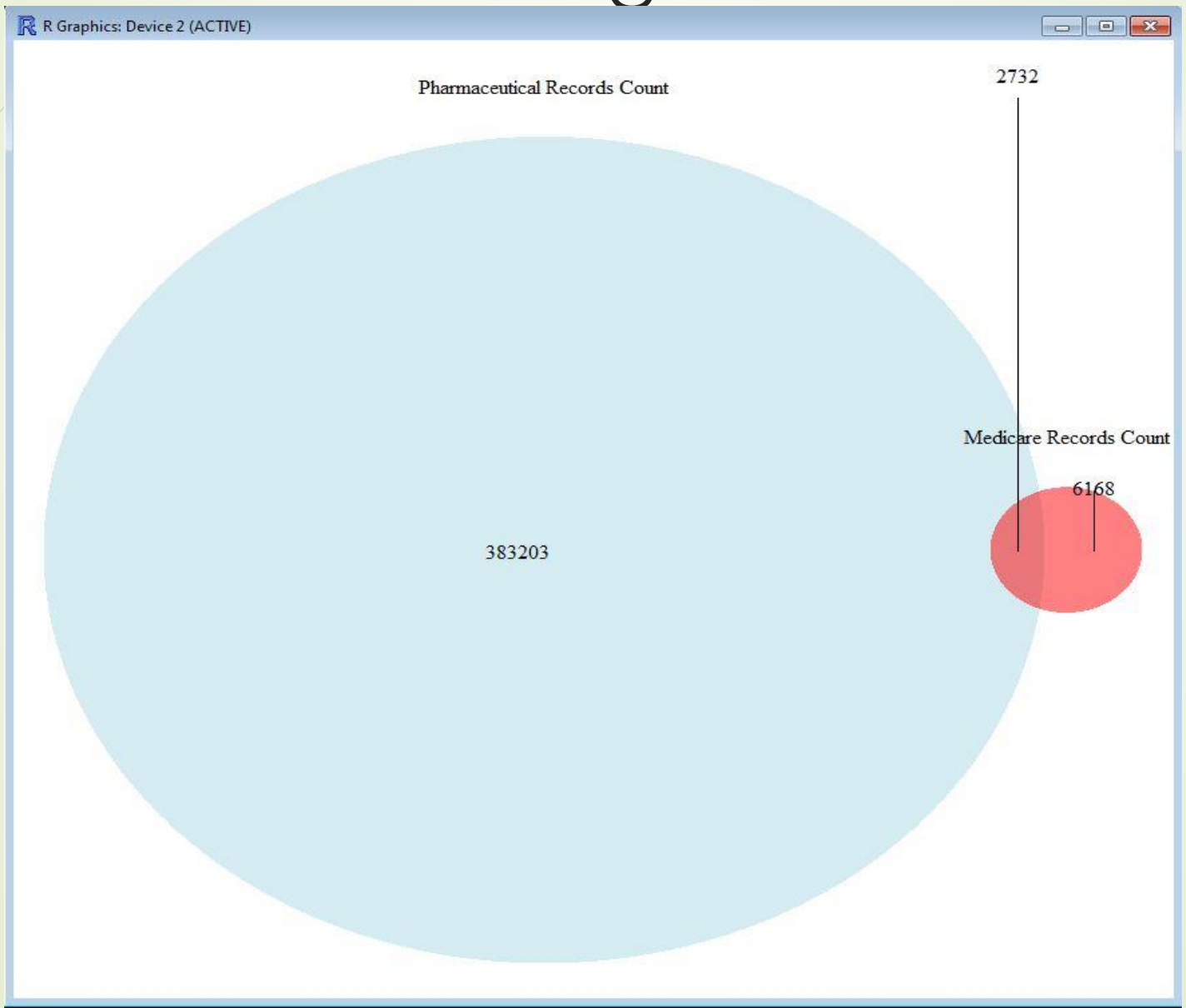
Cardiovascular Stress Test with CPT Code 93015

Correlation formula

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$



Venn diagram



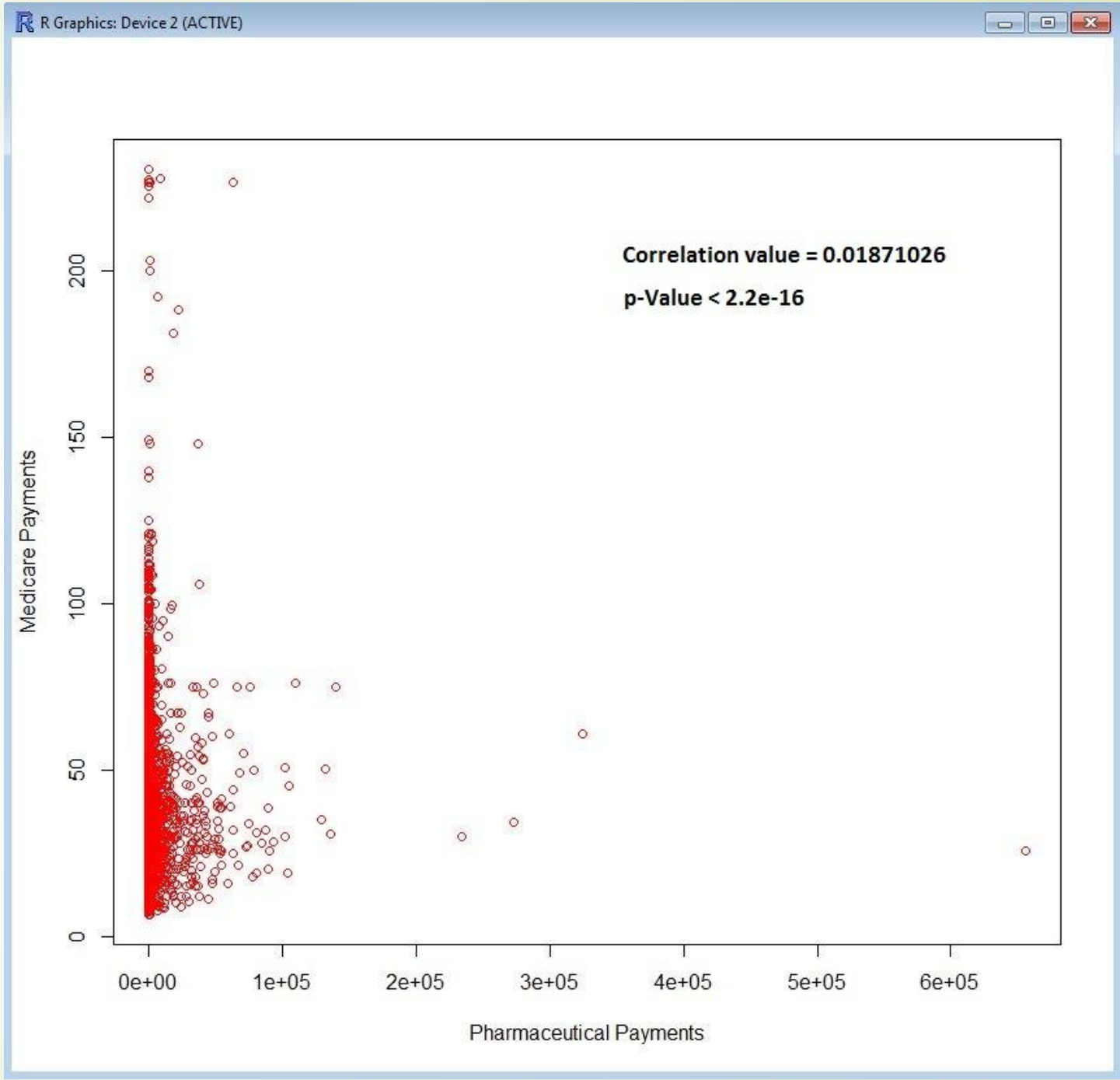
Cardiovascular Stress Test with CPT Code 93015

Extra credit: Scatter plot

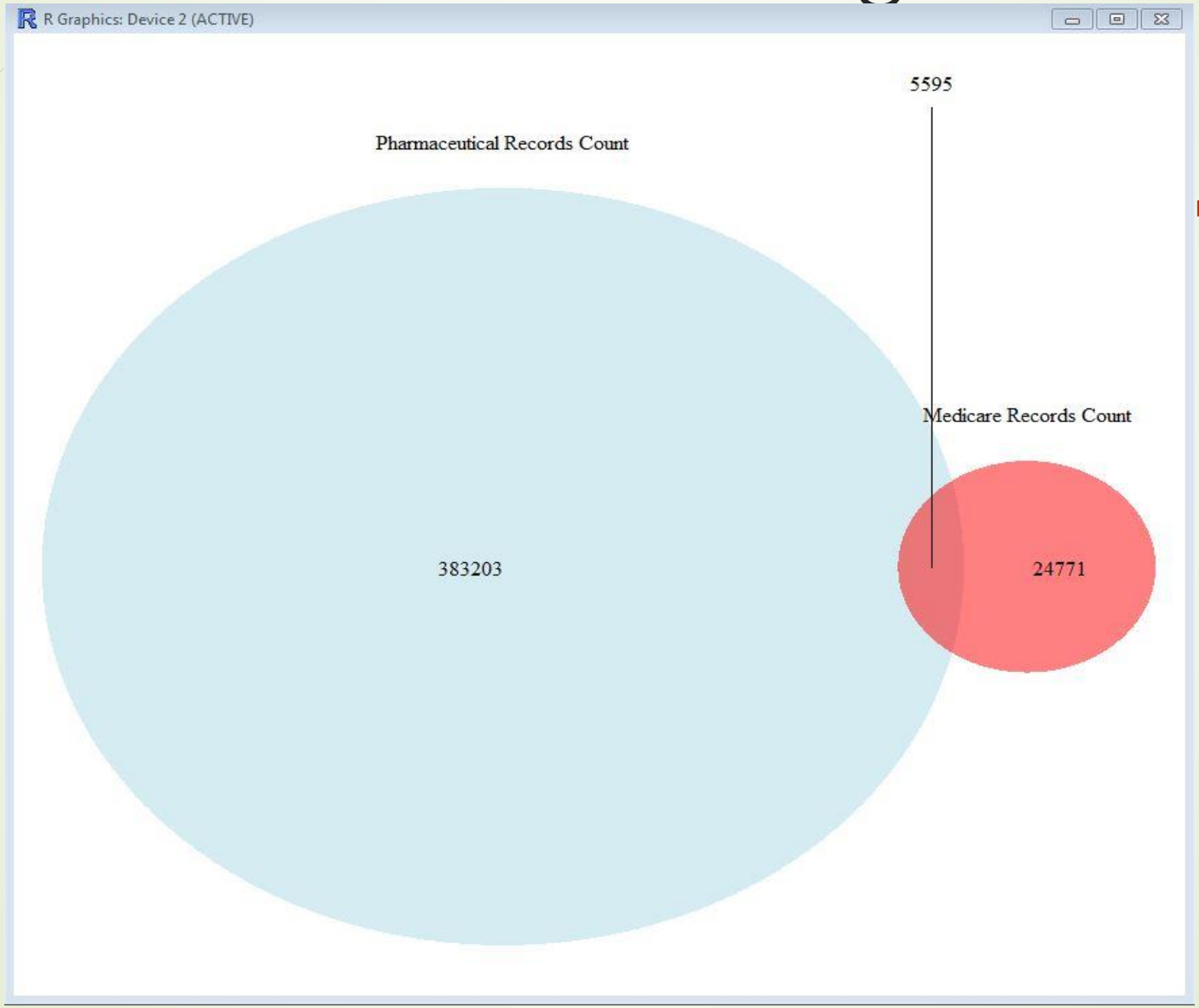
- Electrocardiogram with CPT code 93010

Correlation Formula

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$



Extra credit: Venn diagram



- Electrocardiogram with CPT code 93010

Conclusion

- Thus from the scatter plot and correlation value of 0.003328825, we can conclude that the average Medicare billings in the case of the "Cardiovascular stress test" and total pharmaceutical costs are not closely related to each other.

- The scatter plot, Venn diagram, and correlation value of 0.01871026 for "Electrocardiogram" with CPT code 93010 also shows that there is no real correlation between the average Medicare billings and pharmaceutical payments