



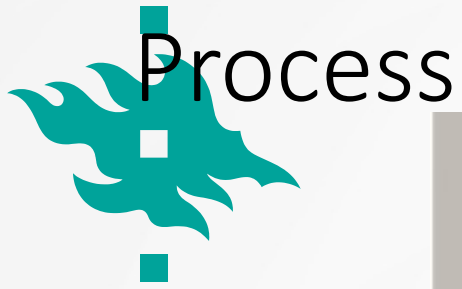
Introduction to Health Facility Business Intelligence

WORKSHOP 3 / Analytics

Jari Haukka, PhD, university lecturer

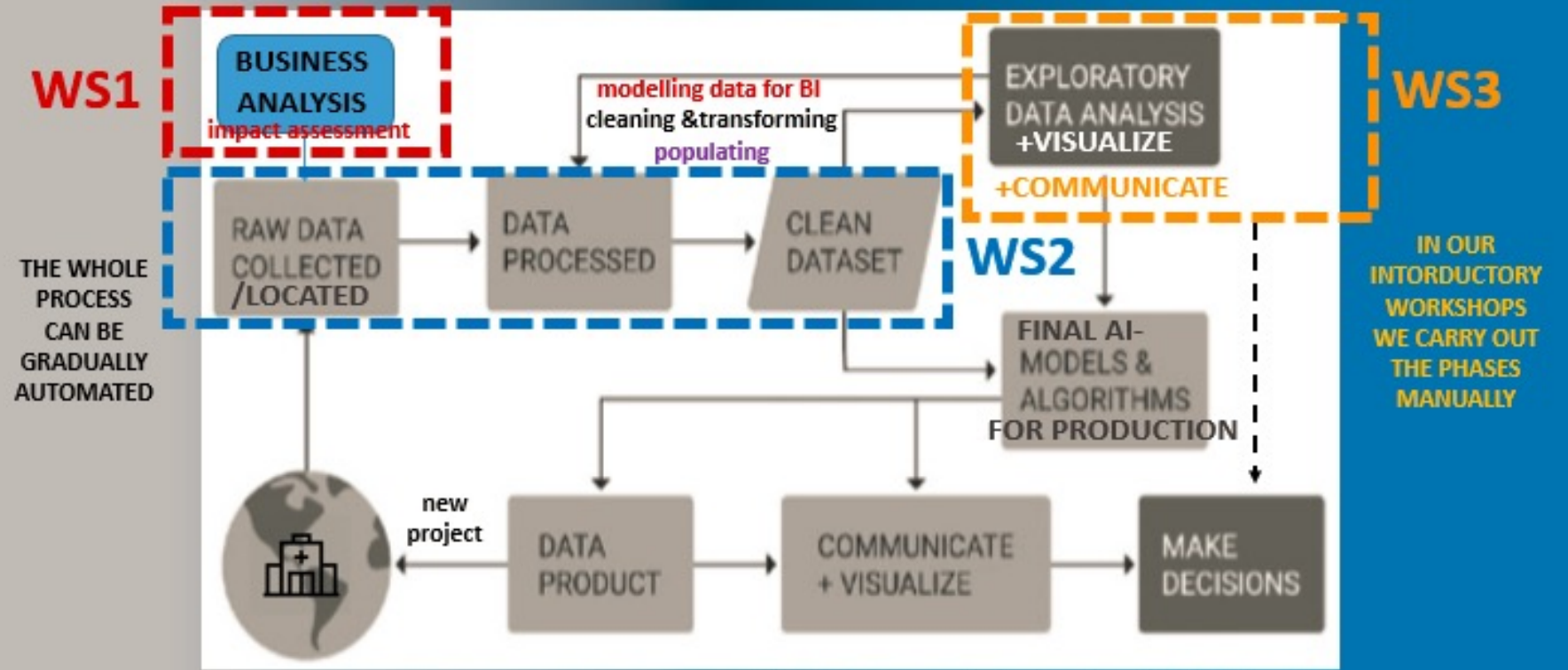


- 10:40 – 10:55 Uses of analytics, models. scenarios and simulations + discussion (Dr. Jari Haukka)
- 10:55 – 11:30 Application to health facility data and KPIs (Dr. Jari Haukka)
- 10:30 – 10:40 Technical platform and tools (Dr. Jari Haukka)
- 10:55 – 11:10 Presentation of exercises (Dr. Jari Haukka)



DATA SCIENCE PROCESS

(MAIN LEARNING GOAL OF THE WS-SERIES!)



Modified from Panesar, Machine Learning and AI for Healthcare, 2019

4.8.2021

Dr. Auvo Finne

Goal of data-analyses

Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. *Journal of Big Data* 2019; 6: 54.

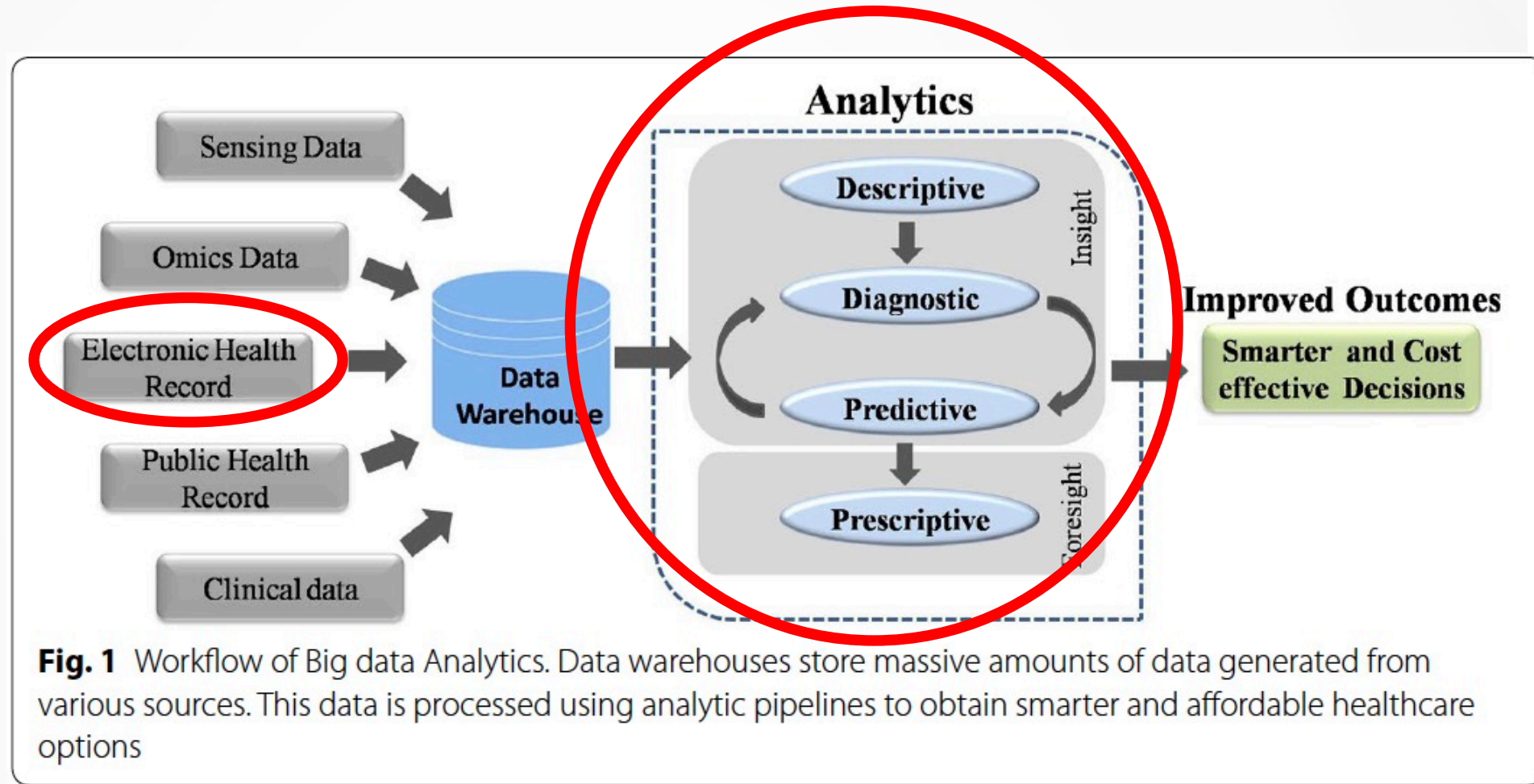


Fig. 1 Workflow of Big data Analytics. Data warehouses store massive amounts of data generated from various sources. This data is processed using analytic pipelines to obtain smarter and affordable healthcare options



"Some say you only get what you measure. I say that's all you get." — A. Donald Stratton

- It is not enough to simply create a numeric measure.
- The measure should accurately reflect the process. We use metrics to base decisions on and to focus our actions.
- It is not only important to measure the **right indicators**, it is important to **measure them well**.
- Choosing the right metrics is critical to success.
- Although there may never be a single perfect measure, it is certainly possible to create a measure or even multiple measures which reflect the performance of your system.
- If the metrics are chosen carefully, then, in the process of achieving their metrics, managers and employees will make the right decisions and take the right actions that enable the organization to maximize its performance.
- <http://www.aleanjourney.com/2014/03/lean-quote-you-get-what-you-measure.html>



Goal of data-analyses

- Smarter decisions
- More cost effective decision
- Open: <https://presemio.helsinki.fi/hfbi> and answer the questions
- Choosing right key performance indicators (KPI) is crucial
- No data analyses are useful, if KPI used are not relevant and well measured
- Here we concentrate on examples on data analyses with some predefined indicators



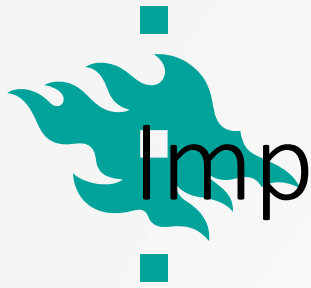
Process of data-analyses

- Extracting data from database
- Importing into analyse software
- Exploratory data analyses (EDA)
 - Basic distributions
 - Tables
 - Graphs, visualization
- Modelling, data analyses
 - Regression models
 - More advanced data science tools



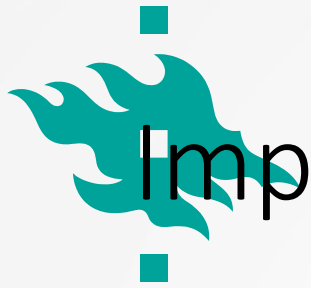
Extracting data from database

- Define relevant variables
- Use scripts to update data extraction
- You may want daily, weekly, monthly, or annuala update of data
- Choose format for data
 - CVS
 - EXCEL
 - Etc.



Importing into analyse software

- Prefer data analyses tools that support scripting in order to smooth import for updating
- After importing data check them carefully



Importing into analyse software/Example

























- We use real data from Health Data New York (<https://data.world/healthdatany>)
 - Hospital Inpatient Discharges (SPARCS De-Identified)
- The Statewide Planning and Research Cooperative System (SPARCS) Inpatient De-identified downloadable file contains discharge level detail on patient
 - Characteristics
 - diagnoses,
 - Treatments
 - Services
 - charges.



Example/ Variables









NYSDOH_HospitalInpatientDischarges_SPARCS_De-Identified_2012.csv

[Request more info](#)

COLUMN NAME	TYPE	DESCRIPTION
 hospital_service_area 	string	
 hospital_county 	string	
# operating_certificate_number 	integer	
# permanent_facility_id 	integer	
 facility_name 	string	
 age_group 	string	
 zip_code_3_digits 	string	
 gender 	string	
 race 	string	
 ethnicity 	string	
 length_of_stay 	string	
 type_of_admission 	string	
 patient_disposition 	string	



Example/ Variables

 discharge_year 	year
# ccs_diagnosis_code 	integer
 ccs_diagnosis_description 	string
# ccs_procedure_code 	integer
 ccs_procedure_description 	string
# apr_drg_code 	integer
 apr_drg_description 	string
# apr_mdc_code 	integer
 apr_mdc_description 	string
# apr_severity_of_illness_code 	integer
 apr_severity_of_illness_description 	string
 apr_risk_of_mortality 	string
 apr_medical_surgical_description 	string



Example/ Variables

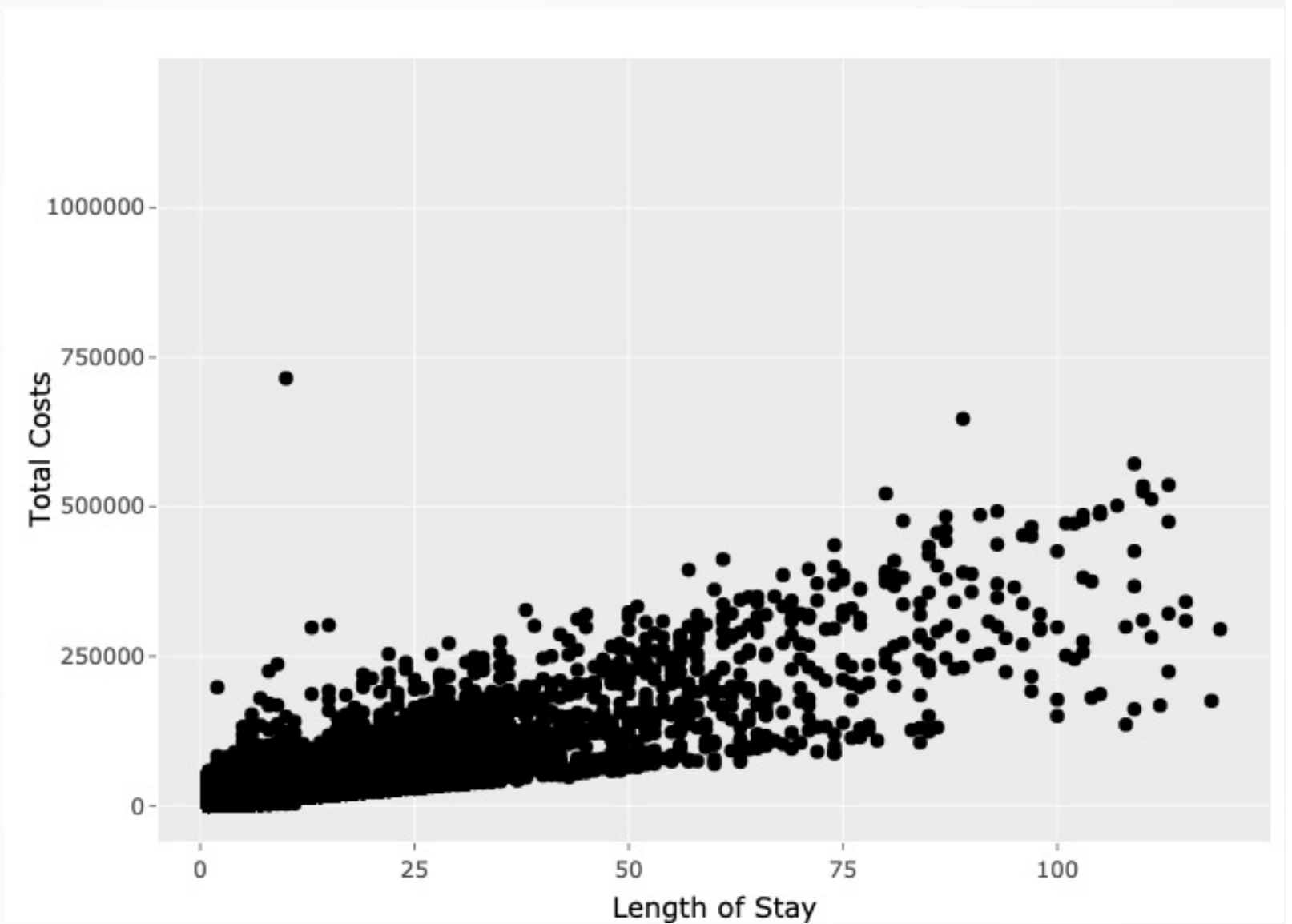
<input type="checkbox"/>	payment_typology_1	i	string
<input type="checkbox"/>	payment_typology_2	i	string
<input type="checkbox"/>	payment_typology_3	i	string
#	birth_weight	i	integer
<input type="checkbox"/>	abortion_edit_indicator	i	string
<input type="checkbox"/>	emergency_department_indicator	i	string
#	total_charges	i	decimal
#	total_costs	i	decimal
#	ratio_of_total_costs_to_total_charges	i	decimal



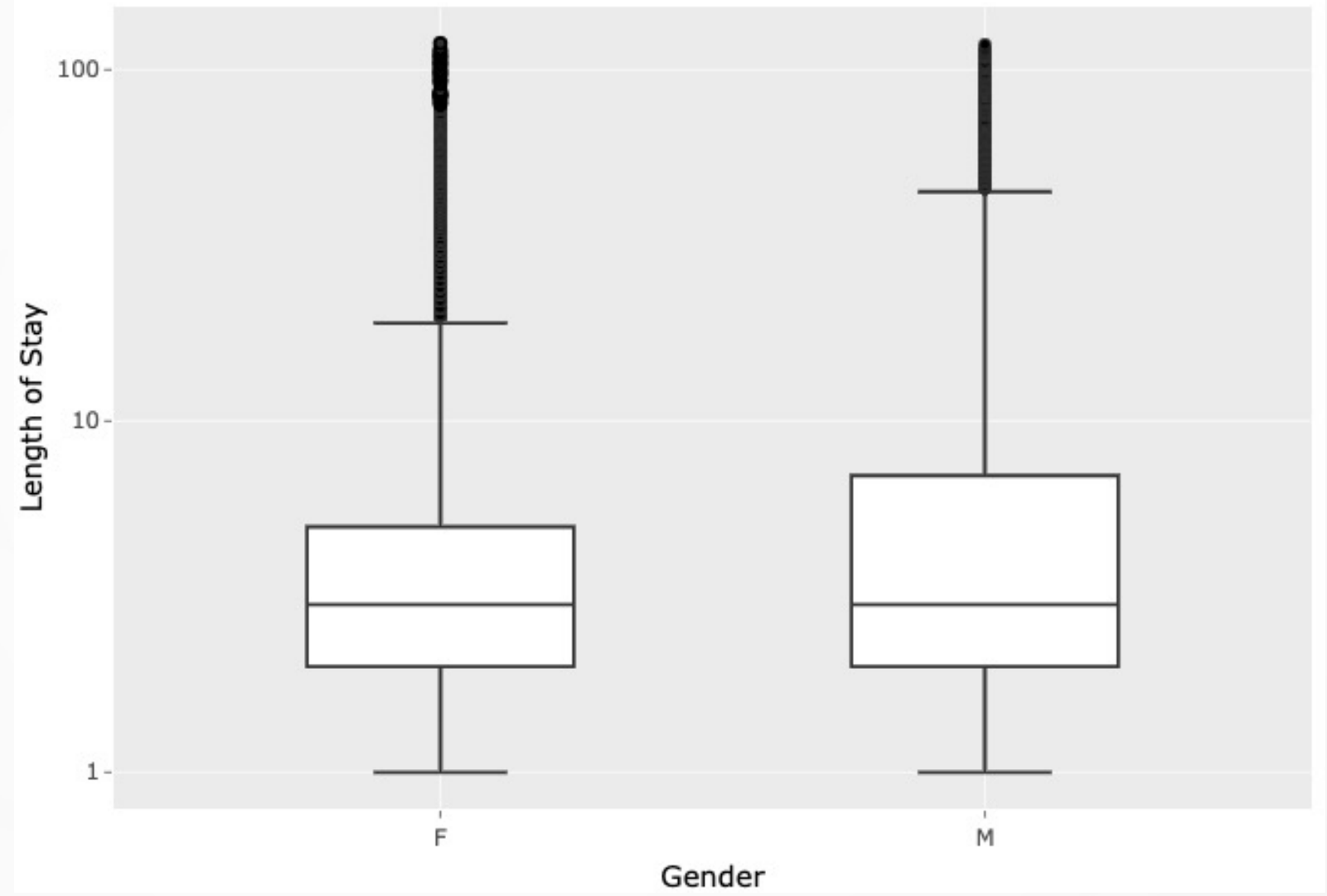
EDA example

- We use R ([R: The R Project for Statistical Computing \(r-project.org\)](https://www.r-project.org/)) and
- Rstudio interface to R ([RStudio | Open source & professional software for data science teams – Rstudio, https://www.rstudio.com/](https://www.rstudio.com/)) software
- Free
- Probably the widest used data science and statistical software

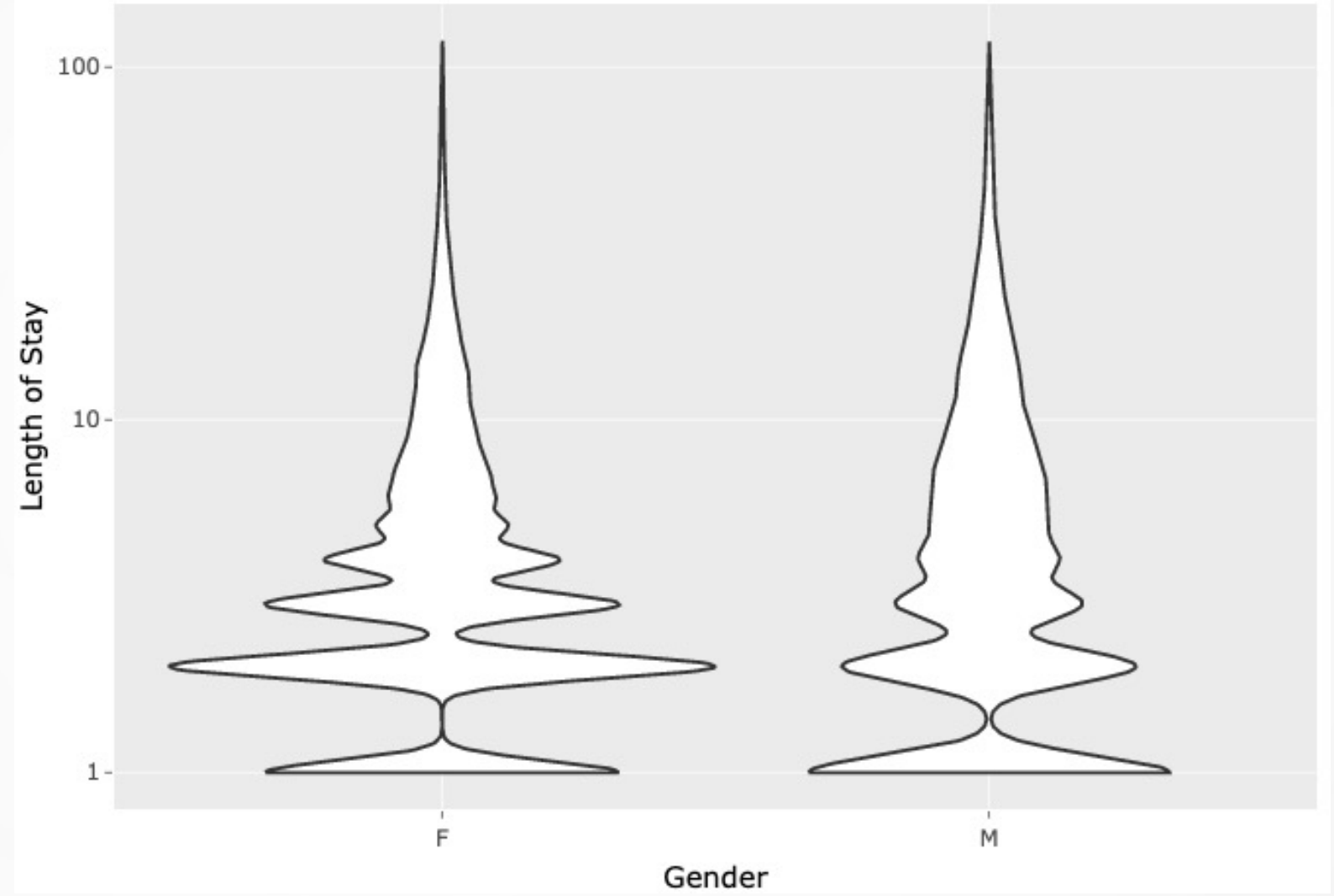
EDA example/ Scatterplot



EDA example/ Boxplot

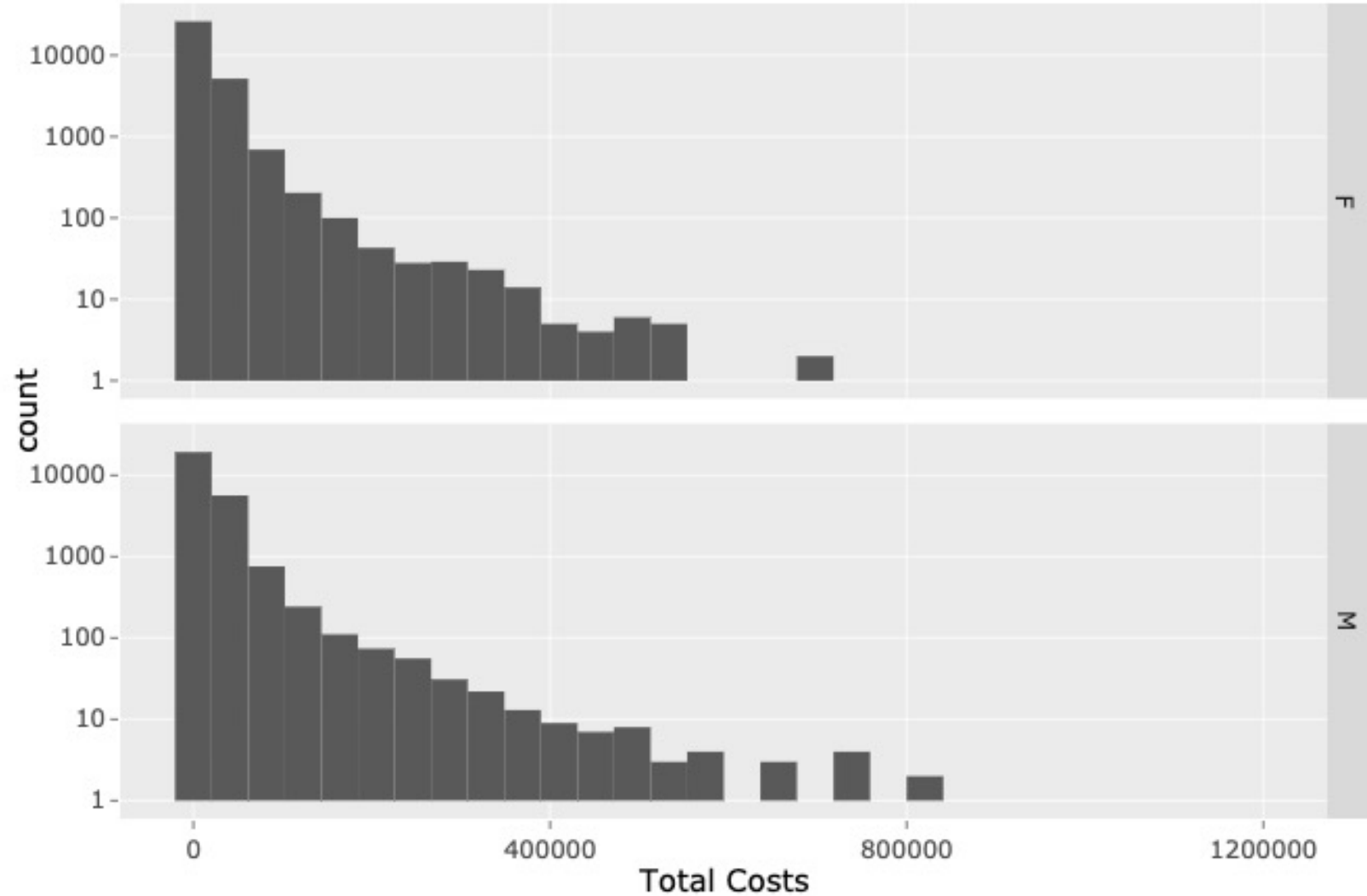


EDA example/ Violin plot



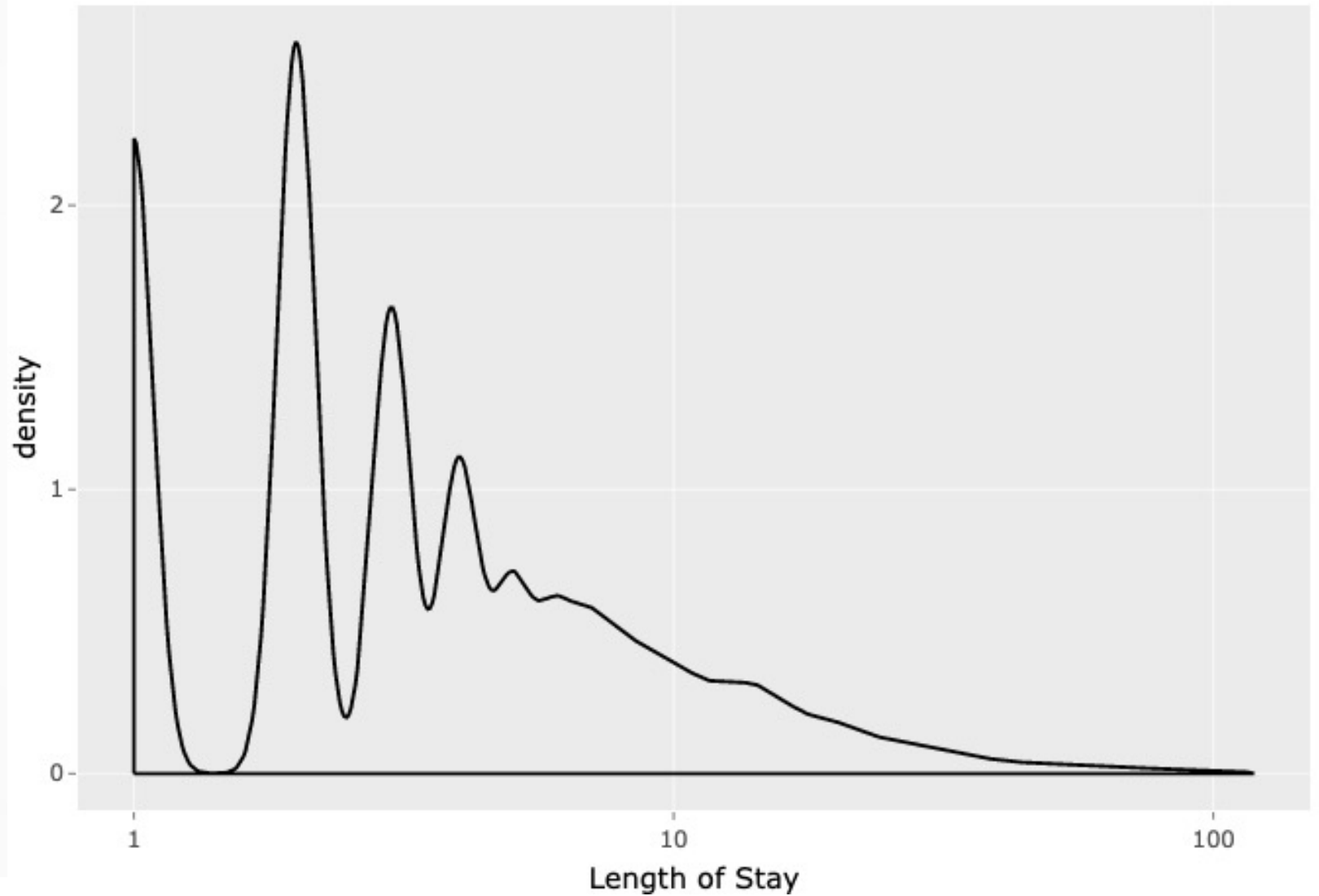
EDA example/ Histogram


NOTE! Log10 x-axis



EDA example/ Density plot

NOTE! Log10 x-axis





EDA example/ Tables

Tables by gender

	F	M	p	test	Missing
n	32438	26284			
Age Group (%)			<0.001		0.0
0 to 17	4701 (14.5)	5294 (20.1)			
18 to 29	3950 (12.2)	1422 (5.4)			
30 to 49	8782 (27.1)	3915 (14.9)			
50 to 69	7636 (23.5)	9454 (36.0)			
70 or Older	7369 (22.7)	6199 (23.6)			
Gender = M (%)	0 (0.0)	26284 (100.0)	<0.001		0.0
Payment Typology 1 (%)			<0.001		0.0
Blue Cross/Blue Shield	3445 (10.6)	2922 (11.1)			
Federal/State/Local/VA	23 (0.1)	8 (0.0)			
Managed Care, Unspecified	181 (0.6)	179 (0.7)			
Medicaid	8069 (24.9)	6574 (25.0)			
Medicare	10497 (32.4)	9778 (37.2)			
Miscellaneous/Other	152 (0.5)	275 (1.0)			
Private Health Insurance	9714 (29.9)	6047 (23.0)			
Self-Pay	357 (1.1)	501 (1.9)			
Emergency Department Indicator = Y (%)	10885 (33.6)	10130 (38.5)	<0.001		0.0
Length of Stay (mean (SD))	5.19 (7.64)	5.91 (9.03)	<0.001		0.1



Regression analyses

- In linear regression we have the following model:

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \dots + x_{ip}\beta_p = \sum_{j=1}^p x_{ij}\beta_j + \epsilon_i, i=1\dots N$$

- x : predicting (independent) variables
- y : response (dependent) variable
- β : coefficients
- p : number of x -variables
- N : number of observations
- ϵ_i : residual ($\epsilon \sim N(0, \sigma_{error})$).



Regression analyses/Example

- Y: `Length of Stay` (response or dependent variable)
- x: predicting (independent) variables
 - `Age Group`
 - Gender
 - `Type of Admission`,

- R code:

```
glm(formula = `Length of Stay` ~ `Age Group` + Gender +  
`Type of Admission`, ## data = eHeatZ.NYSDOH.1)
```



Regression analyses/Example

```
## Call:
## glm(formula = `Length of Stay` ~ `Age Group` + Gender + `Type of Admission`,
##      data = eHeatZ.NYSDOH.1)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -9.311  -3.109  -1.684   0.389  112.504
##
## Coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.30809    0.15325  21.586 < 2e-16 ***
## `Age Group`18 to 29    0.23603    0.17999   1.311  0.18976
## `Age Group`30 to 49    0.25770    0.15942   1.616  0.10600
## `Age Group`50 to 69    0.48306    0.15340   3.149  0.00164 **
## `Age Group`70 or Older  0.68627    0.15643   4.387 1.15e-05 ***
## GenderM                0.31750    0.06871   4.621 3.82e-06 ***
## `Type of Admission`Emergency  2.50193    0.07786  32.133 < 2e-16 ***
## `Type of Admission`Newborn    0.37550    0.17859   2.103  0.03551 *
## `Type of Admission`Not Available  4.72786    0.42476  11.131 < 2e-16 ***
## `Type of Admission`Urgent     5.99921    0.11213  53.504 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Regression analyses/Example/Interpretation

1. There no difference between unde 18 and 18 to 29, or 30 to 49
2. Age Group 50 to 69 has 0.48 d longer stay compared to under 18
3. Age Group 70 or older has 0.68 d longer stay compered to under 18

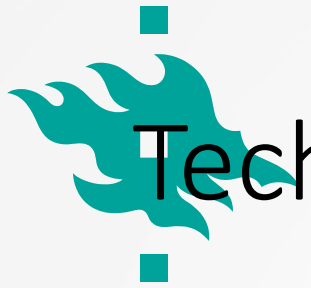
```
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   3.30809    0.15325   21.586 < 2e-16 ***
## `Age Group` 18 to 29          0.23603    0.17999    1.311  0.18976
## `Age Group` 30 to 49          0.25770    0.15942    1.616  0.10600
## `Age Group` 50 to 69          0.48306    0.15340    3.149  0.00164 **
## `Age Group` 70 or Older       0.68627    0.15643    4.387  1.15e-05 ***
```



Regression analyses/Example/Interpretation

1. Males has 0.32 d longer stay compared to females
2. "Emergency" admission has 2.50 d longer stay compared to "Elective" and "Urgent" 5.99 d longer
3. There is no difference between "Elective" and "Newborn"

## GenderM	0.31750	0.06871	4.621	3.82e-06	***
## `Type of Admission`Emergency	2.50193	0.07786	32.133	< 2e-16	***
## `Type of Admission`Newborn	0.37550	0.17859	2.103	0.03551	*
## `Type of Admission`Not Available	4.72786	0.42476	11.131	< 2e-16	***
## `Type of Admission`Urgent	5.99921	0.11213	53.504	< 2e-16	***



Technical platform and tools

- We use R ([R: The R Project for Statistical Computing \(r-project.org\)](https://www.r-project.org/))
 - Rstudio interface to R ([RStudio | Open source & professional software for data science teams – Rstudio, https://www.rstudio.com/](https://www.rstudio.com/)) software
- [Project Jupyter | Home \(https://jupyter.org/\)](https://jupyter.org/)
 - develops open-source software, open-standards, and services for interactive computing across dozens of programming languages (R, Python)
- Other tools:
 - SAS
 - SPSS
 - EXCEL
 - Etc.



Presentation of exercises

1. Read carefully data analyses in Moodle (html, pdf, and doc are identical) (tables, figures, regression analyses) (also in <https://jari-haukka.shinyapps.io/HFBI/>)
2. Write description of data using tables and figures. In this assignment we use hospital stay length as key performance indicator (KPI). You may use copy-paste to include tables and figures in you report.
3. Comment briefly raw data. Is there any interesting features?
4. Interpret both regression regression models below. Did adding variable Payment Typology 1 change interpretation of model?