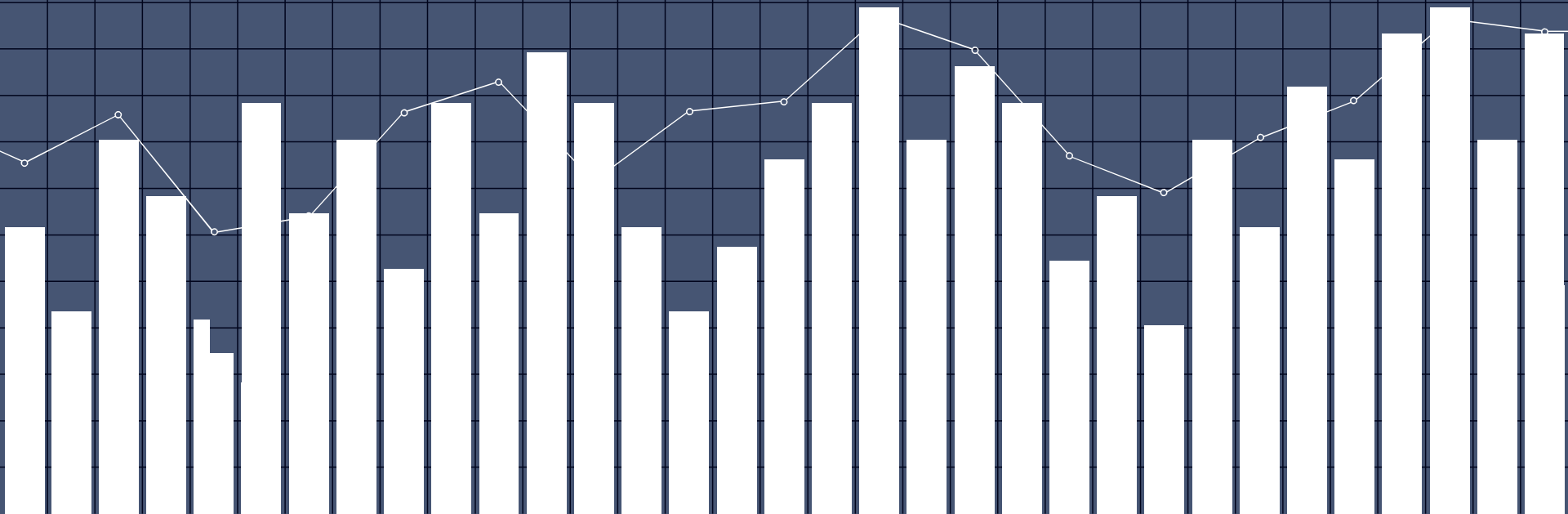
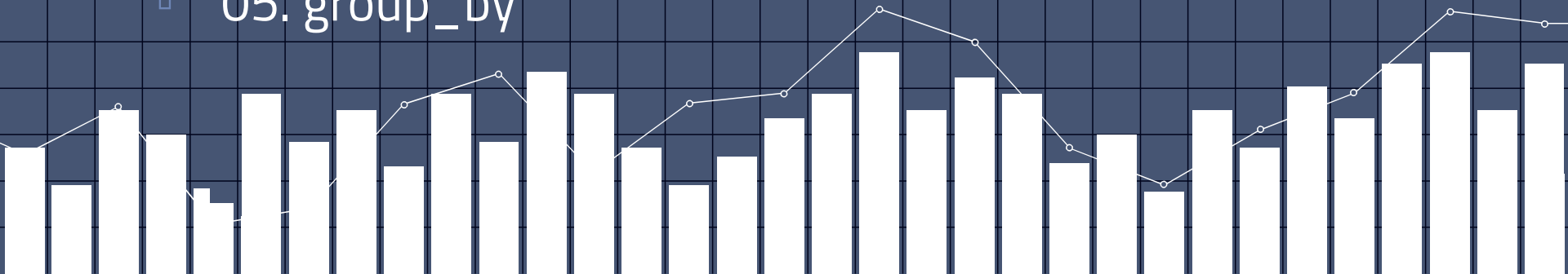


05. Manipulace s daty



Harmonogram

- 01. tbl
- 02. select, mutate
- 03. filter, arrange
- 04. summarise, %>%
- 05. group_by



dplyr a tibble

```
# Načtení knihovny  
library(dplyr) # install.packages("dplyr")
```

```
# Data  
Titanic = read.csv("Train.csv")
```

```
# Typ objektu  
class(Titanic)
```

```
# konverze na tibble  
Titanic = as_tibble(Titanic)
```

```
# Struktura dat  
glimpse(Titanic)
```

```
# Pojmenování proměnných  
Port = c("C" = "Cherbourg", "Q" = "Queenstown",  
"S" = "Southampton")
```

```
Titanic$Embarked =  
as.character(Titanic$Embarked)
```

```
Titanic$Embarked = Port[Titanic$Embarked]
```

```
# Struktura dat  
glimpse(Titanic$Embarked)
```

**DON'T WORRY ABOUT THE LIFEBOATS
THEY SAID**



THIS SHIP IS UNSINKABLE THEY SAID

select

Manipulace s proměnnými

Výběr proměnné výčtem

```
select(Titanic, Name)
```

Výběr více proměnných - pořadím

```
select(Titanic, PassengerId:Survived, Name:Age)
```

Výběr více proměnných – na základě znění

```
select(Titanic, contains("e"))
```

Výběr více proměnných – kombinace postupů

```
select(Titanic, 1:2, contains("t"))
```

mutate

Manipulace s proměnnými

Tvorba nových proměnných

```
Titanic_2018 = mutate(Titanic, Age_2018 = (Age + (2018-1912)))
```

Tvorba nových proměnných – více případů současně

```
Titanic_02 = mutate(Titanic, Price_Year_Ratio = (Fare / Age), Family = Sibsp + Parch)
```



Sorry Sir. Women and children first.



Did you just assume my gender?

filter

Manipulace s pozorováními

Výběr na základě logických operátorů

```
filter(Titanic, Age < 18)
```

```
filter(Titanic, Embarked %in% c("Southampton", "Queenstown"))
```

Kombinace operátorů

```
filter(Titanic, Age < 18, Survived == 1)
```

```
filter(Titanic, Age < 18, Pclass == 1 | Pclass == 2)
```

Filtrace na základě textu

```
library(stringr)
```

```
filter(Titanic, str_detect(Name, str_detect(Name, "Rose"))
```


arrange

Manipulace s pozorováními

```
# Seřazení přeživších podle věku, bez chybějících hodnot  
Survived = filter(Titanic, Survived == 1, !is.na(Age))  
arrange(Survived, Age)
```

```
# Seřazení přeživších podle věku od nejstarších po nejmladší, bez chybějících hodnot  
arrange(Survived, desc(Age))
```

```
# Seřazení přeživších podle věku od nejstarších po nejmladší a místa nalodění  
arrange(Survived, desc(Age), Embarked)
```

summarize

Agregační funkce

Nejnižší cena za lístek a nejstarší osoba

```
summarize(Titanic, Fare_Min = min(Fare), Age_Max = max(Age))
```

Průměrný věk těch, kdo přežili

```
summarize(filter(Titanic, Survived == 1), Age_Avg = mean(Age, na.rm = TRUE))
```

Průměrná cena palubního lístku a její směrodatná odchylka, median, minimum, maximum

```
Titanic_03 = filter(Titanic, !is.na(Fare))
```

```
summarize(Titanic_03, Mean_Fare = mean(Fare), SD_Fare = sd(Fare), Median_Fare =  
median(Fare), Min_Fare = min(Fare), Max_Fare = max(Fare))
```

%>%

Pipe operator

Průměrný věk osob z 3. třídy, které nepřežily

Titanic %>%

```
filter(Pclass == 3) %>%
```

```
filter(Survived == 0) %>%
```

```
summarize(Age_Died_Avg = mean(Age, na.rm = TRUE))
```

Nejvyšší cena do 1. třídy pro osobu, která se jmenuje William a která přežila

Titanic %>%

```
filter(Pclass == 1) %>%
```

```
filter(str_detect(Name, "William")) %>%
```

```
filter(Survived == 1) %>%
```

```
summarize(Fare_Avg_Will = min(Fare))
```



CECI N'EST PAS UNE %>%.

group_by

Pipe operator

Průměrný věk osob z 3. třídy, které nepřežily

```
Titanic %>%
```

```
  group_by(Sex) %>%
```

```
  summarize(Mean_Fare = mean(Fare, na.rm = TRUE), Count = n(), Mean_Age =  
            mean(Age, na.rm = TRUE))
```

Zdroje

Data Wrangling with dplyr and tidyr Cheat Sheet – <https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf>

