Deel 2: Data Representatie

Aan de hand van onderstaande vragen, bestudeer je de titanic dataset.

Tip: sommige woorden hieronder zijn bold gedrukt, dit kan een hint zijn naar de oplossing

1. Maak een Jupyter Notebook met als naam: *jouwnaam_jouwvoornaam_titanic*
   Alle oplossingen noteer je in dit Jupyter Notebook.

2. Download titanic.csv van BB en plaats deze in de map waar je ook bovenstaande Jupyter Notebook gezet hebt.

3. Mbv de **read_csv()** methode uit de Pandas library in Python, laad je deze csv – file in jouw Jupyter notebook. Geef deze de naam *'test_titanic'*.
   Hiervoor moet je uiteraard eerst de Pandas library importeren in jouw notebook.
   Geef deze import de naam "pd". Dit is handig om in het vervolg mee te werken.

   ```
   import pandas as pd
   ```

   Hulp?
   PluralSight: Pandas Playbook: Visualisation
   3) Making Simple Plots
   ➔ Demo: Introducing the Pandas Plot API

4. Een (ietwat andere) titanic dataset zit ook vervat in de Seaborn library van Python.
   De verdere analyses in dit document voer je uit op deze titanic – dataset.
   Laad deze dataset in jouw Jupyter Notebook. Uiteraard moet je hiervoor eerst de Seaborn library importeren. Geef deze de naam "sns".
   Dit kan je doen door volgende 2 commando's

   ```
   import seaborn as sns
   titanic = sns.load_dataset('titanic')
   ```

   Merk op: Opgave 3 is wel belangrijk: op het examen moet je kunnen werken met zowel excel / csv – bestanden die wij voor jullie voorzien en in jouw notebook kunnen inladen alsook datasets die vervat zitten in een library van Python.

5. Om een zicht te krijgen op de inhoud van jouw dataset, print je de eerste 5 (**hoofd** (methode head()) van dataset) en laatste 5 (**staart** (methode tail())van dataset) rijen van deze dataset eens uit.
   Hulp?
   PluralSight: Cleaning Data: Python Data Playbook
   2) Understanding Your data
   ➔ Viewing and converting types

6. Welke variabelen (features = kolommen) zijn opgenomen in deze dataset?
   Maak een afdruk van de kolommen samen met het type (field **dtypes**) van de feature.
   Hulp?
   PluralSight: Cleaning Data: Python Data Playbook
   2) Understanding Your data
   ➔ Viewing and converting types

7. Je kan in Python allerhande informatie bekomen uit een dataset door gebruik te maken van de methode **info()** bvb het aantal passagiers (rijen dataset)**.** Hoeveel passagiers (rijen) hebben we in deze dataset?

Korte herhaling: werken met datasets in Python

Om een kolom (feature) te selecteren kan je de " . " + "naam kolom" gebruiken of je werkt met [ ] en hiertussen plaats je de naam van de kolommen die je wil selecteren.
Merk op: selecteren van de feature "class" uit de titanic dataset werkt enkel met [ ].
Reden? Class is een gereserveerd woord binnen Python

8. Selecteer de kolom 'class' uit de dataset.
   Selecteer 3 kolommen 'embark_town, 'alive' en 'alone' en print de gegevens af voor alle passagiers.
   Hulp?
   PluralSight: Exploratory Data Analysis with Python
   7) Practicing Data Analysis with Python
   ➔ 04 Take a look at the data

Om één rij (of meerdere rijen) te selecteren, geef je tussen [ ] aan welke rijen je nodig hebt.

9. Selecteer passagier 5 tot 10

   Hulp?
   PluralSight: Python for Data Analysis
   3) Leveraging Built-in Functions and Complex Data Types
   ➔ Demo: introducing List
   ➔ Demo: List Slicing Operations

   Hulp?
   docs.python.com
   ➔ tutorial
   3.1.3 Lists

10. Selecteer voor passagier 5 tot 10 de kolommen 'survived' en 'age' en print af.

11. Hoeveel mannen bevat deze dataset?
    a. Bereken dit aantal door een selectie te doen op geslacht: je creëert een nieuwe dataset (enkel mannen) en vraagt de **lengte** van deze set op
    b. Er bestaan ook een voorgeprogrammeerde functies (bvb **value_counts()** die je kan toepassen op de feature 'sex' (je krijgt dan zowel het aantal mannen als het aantal vrouwen)

12. Hoeveel mannen reisden in eerste klasse? Bereken dit aantal via een selectie.

13. Bepaal het percentage mannen / vrouwen in deze dataset (je kan dit bvb doen door een parameter **(normalize = True)** toe te voegen aan de functie die je gebruikt hebt in 11: **value.counts()**)

14. Bereken het aantal personen per embark_town.
    Is voor elke passagier de feature 'embark_town' ingevuld? Druk de rijen waarbij de feature 'embark_town' een null-waarde is, af. Gebruik hiervoor de **isnull()** methode.Druk nu het aantal lege rijen per feature af (Tip: Maak de **som** van het aantal rijen die "**isnull**" zijn).

15. Wat doet de functie **describe()?**

16. Wat gebeurt er indien je in de functie **describe()** de parameter **include = 'all'** toevoegt?

17. Bereken telkens met een aparte functie voor de observatie 'fare':
    a. Het gemiddelde (mean())
    b. De mediaan (median())
    c. Het minimum (min())
    d. Het maximum (max())
    e. De variantie (var())
    f. De standaardafwijking (std())
    g. Het eerste kwartiel (quantile(.25))
    h. Het tweede kwartiel (quantile(.5))
    i. Het derde kwartiel (quantile(.75))

    Hulp?

    PluralSight: Interpreting Data Using Descriptive Statistics with Python
    3) Working with Descriptive Statistics using Pandas
    4) Working with Descriptive Statistics using Scipy and Statsmodels

    Wat stellen deze grootheden voor?
    Kan je deze grootheden ook berekenen voor alle andere variabelen (bvb pclass, survived…)?

GRAFISCH

Hulp?
Zowel de library Matplotlib als Pandas voorzien in methoden om plots te creëren in Python.

18. Maak een cirkeldiagram **(pie)** van de variabele 'survived'.

19. Maak een cirkeldiagram van de variabele 'survived' per geslacht. Vorm een aparte dataset met mannen resp. vrouwen en maak daarna het cirkeldiagram voor de variabele 'survived'.

20. Maak een staafdiagram **(bar)** van de variabele 'pclass'.

21. Maak een staafdiagram per geslacht van de variabele 'pclass'.

22. Maak een histogram **(hist)** van de variabele 'leeftijd'.

23. Maak een histogram van de variabele 'fare', zorg ervoor dat je 20 **bins** hebt.
    bins = aantal staafjes: zie cursus data representatie pg 74 en 75: juiste keuze klassenbreedte

24. Wat is het verschil tussen een staafdiagram en een histogram?
    Wanneer gebruiken we het ene / andere?

25. Maak een **boxplot** van de variabele 'age'.
    Indien je foutmeldingen krijgt ivm NAN-waarden, neem dan enkel de **notnull()** rijen mee voor het creëren van de boxplot.

26. Maak een boxplot van de variabele 'age' opgesplitst naar pclass.

27. Welke gegevens zijn opgenomen in een boxplot?
    Welke info kan je halen uit een boxplot?

VERBANDEN TUSSEN VARIABELEN

28. Maak een kruistabel (**crosstab van pandas**) van de variabele pclass en geslacht.
    Welke info kan je uit een kruistabel aflezen?
    Voeg de parameter normalize = True eens toe. Welke info lees je nu af?

29. Maak een pivottabel (**pivot_table**) voor de gemiddelde leeftijd per pclass en per geslacht.

30. Maak een scatterplot (**scatter van malplotlib**) van de variabelen “age” en “fare”.

31. Maak een heatmap in python die correlaties weergeeft tussen de features van de titanic – dataset.

MISSING VALUES + OUTLIERS

Mbv opgave 7 en/of 14 kan je afleiden dat er per feature een aantal observaties niet ingevuld zijn (dus missing). Druk deze **info** (nl hoeveel observaties er per feature niet ingevuld(**isnull** zijn)) nog even af.

32. De feature “embarked” heeft voor 2 passagiers geen gegevens. Zoek in de dataset naar deze missing observaties (**isnull**). Print deze 2 rijen af.

33. Welke manieren kan je (online) vinden om met “missing data” om te gaan?

34. Wij kiezen voor “imputation”. Indien we de missing values voor “embarked” zouden opvullen met de plaats waar de meeste mensen opstapten, dan vul je op met … ?
    Tip: vraag 11 kan je hierbij helpen

35. Maar… bekijk de rijen van de missing values voor “embarked” even dieper. We zien dat beide passagiers de ramp overleefden. Zoek uit van welke opstapplaats het meeste passagiers overleefden. Dit kan je bekomen via een kruistabel (**crosstab** in pandas). Via deze manier van redeneren kom je tot plaats …

36. Wanneer we nog verder kijken zien we dat beide passagiers als fare 80 hebben en beiden in eerste klas reisden. Zoek welk opstappunt voor klasse 1 als median waarde voor fare ongeveer 80 heeft. Dan kom je uit bij….
    Tip: we groeperen (**groupby**) de titanic – dataset op basis van “pclass” en “embarked” en berekenen voor elke categorie de mediaan (**median**) van de feature fare”.

37. Vul de 2 missing values van embarked op met C. Er zijn meerdere manieren om dit te doen:
    - selecteer uit de dataset de “null” waarden voor ‘embarked” en geef die de waarde ‘C’
    - pas **fillna** – methode toe op de feature “embarked”, let op met parameter: inplace

38. Hoeveel missing values zijn er voor “age”?
    Dit zijn er te veel om allemaal individueel te overlopen. Print 10 willekeurige rijen uit waarbij de feature “age” missing is. Volgend commando kan je hierbij helpen:
    pd.options.display.max_rows = 10

39. Vooraleer we leeftijd gaan opvullen met bvb gemiddelde leeftijd of mediaan van de leeftijd, gaan we even dieper in op de variabele leeftijd:
    - Bestudeer de variabele leeftijd aan de hand van een histogram.
    - Bereken het gemiddelde, de mediaan, het min en het max van “leeftijd”.
    Doe dit ook eens voor mannen en vrouwen apart.
    - Maak een boxplot van de variabele “leeftijd” per geslacht.
    Conclusie: Is er een “groot” verschil tussen mannen en vrouwen wat betreft leeftijd?

40. We doen hetzelfde maar nu voor pclass. Maak een boxplot per pclass van de variabele leeftijd.

41. Vervang de null waarden voor de variabele leeftijd door de mediaan van de leeftijd per pclass.

42. Er zijn veel missing values voor “deck”. Hoeveel? Deze vullen we niet op.

43. Wat zijn outliers / uitschieters?

44. Maak een histogram voor de feature fare. Zou je hieruit afleiden dat er outliers zijn?

45. Maak een boxplot voor de feature “fare”.

46. Print die rijen af waarbij de feature “fare” gelijk is aan de maximum waarde voor fare. Hoeveel rijen vind je?

47. Welke manieren kan je (online) vinden om met outliers om te gaan?

48. We maken van onze (continue) variabele ”fare” een categorische variabele door deze in 4 categorieën op te delen. Dit kan je doen via de methode **qcut** (uit Pandas)
We geven deze categorieën elk een naam: “zeer laag”, “laag”, “hoog”, “zeer hoog”

49. Zijn de observaties ongeveer gelijkmatig verdeeld over de 4 categorieën? Neem hiervoor de oplossing van vraag 48, pas hier de methode value_counts() en maak er een staafdiagram van.

50. We maken van deze 4 categorieën een extra feature in onze dataset met naam: “fare_category” als volgt:
titanic[‘fare_category’] = … zie oplossing opgave 48

DICHTHEIDSFUNCTIE + NORMALE VERDELING

51. Maak een histogram van de feature ‘fare’

52. Teken de dichtheidsfunctie (**kerneldensityplot; kind = kde)** van de feature ‘fare’. Lijkt jou deze verdeling symmetrisch?

53. Bepaal de ‘scheefheidsparameter’ (**skew**) voor de variabele fare.

54. Maak een histogram van de feature ‘age’.

55. Teken de dichtheidsfunctie van de feature ‘age’. Lijkt jou deze verdeling symmetrisch?

56. Bepaal de ‘scheefheidsparameter’ (**skew**) voor de variabele age.

57. Wij nemen aan (zonder verdere expliciete normaliteitstest) dat de verdeling van de feature ‘age’ normaal is.
    Bereken het gemiddelde en de standaardafwijking van deze feature.

58. Bepaal de kans dat de leeftijd van een passagier tussen 20 en 30 ligt mbv de normale verdeling.

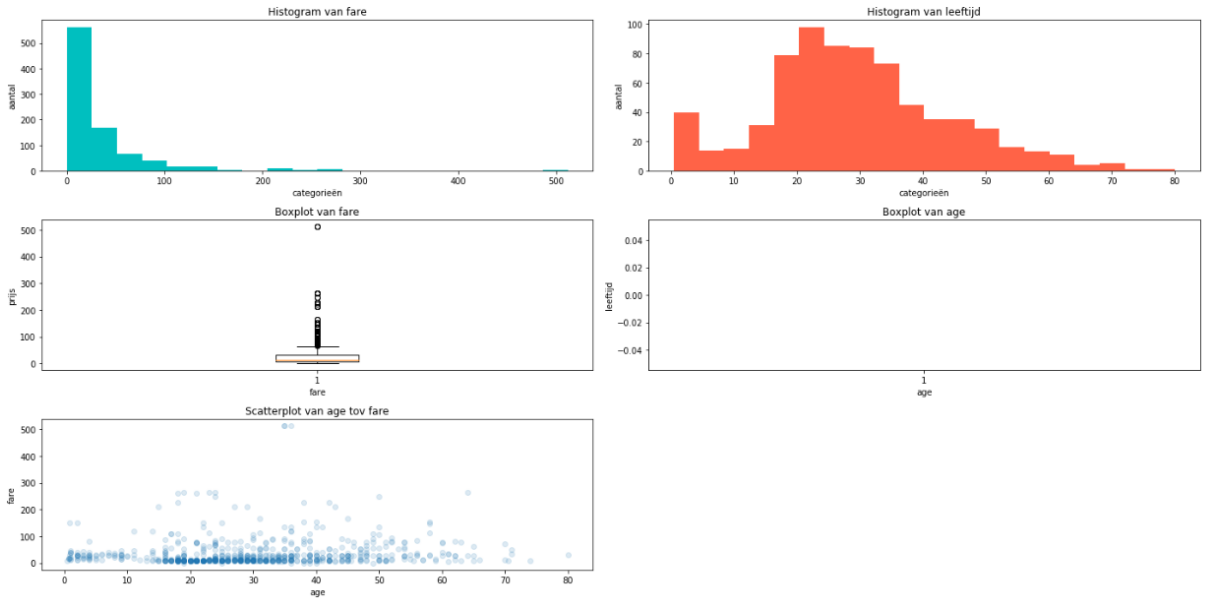
59. Controleer jouw resultaat van opgave 58 met het resultaat dat je bekomt wanneer je gebruik maakt van de steekproef: bepaal exact hoeveel passagiers een leeftijd hebben tussen 20 en 30.

60. Om af te sluiten: geen examenleerstof….
    a. Hoe meerdere subplots tegelijk maken in python?

    b. Hoe kan je deze voorwaardelijke kansen bereken: $P(survived|male)$ en $P(survived|female)$?

Oplossing opdracht 60: a

```
#a: meerdere subplots tegelijk in python
f, ax_array = plt.subplots(3, 2, figsize=(20,10)) # 3 rijen en 2 kolommen

ax_array[0,0].hist(titanic.fare, bins = 20, color = 'c')
ax_array[0,0].set_title('Histogram van fare')
ax_array[0,0].set_xlabel('categorieën')
ax_array[0,0].set_ylabel('aantal')

ax_array[0,1].hist(titanic.age, bins = 20, color = 'tomato')
ax_array[0,1].set_title('Histogram van leeftijd')
ax_array[0,1].set_xlabel('categorieën')
ax_array[0,1].set_ylabel('aantal')

ax_array[1,0].boxplot(titanic.fare.values)
ax_array[1,0].set_title('Boxplot van fare')
ax_array[1,0].set_xlabel('fare')
ax_array[1,0].set_ylabel('prijs')

ax_array[1,1].boxplot(titanic.age.values)
ax_array[1,1].set_title('Boxplot van age')
ax_array[1,1].set_xlabel('age')
ax_array[1,1].set_ylabel('leeftijd')


ax_array[2,0].scatter(titanic.age, titanic.fare, alpha=0.15)
ax_array[2,0].set_title('Scatterplot van age tov fare')
ax_array[2,0].set_xlabel('age')
ax_array[2,0].set_ylabel('fare')

#afstand tussen plots
plt.tight_layout()
#lege subplot onderaan weg (3 bij 2 = 6 plaatsen)
ax_array[2,1].axis('off')

plt.show()
```



Oplossing opdracht 60 b:

```
#c: voorwaardelijke kans
fig, axs = plt.subplots(1, 2, figsize=(10,4))
for i, sex in enumerate(['female', 'male']):
    p = titanic[titanic['sex'] == sex]['survived'].value_counts(normalize=True)
    p.plot(kind='bar', ax=axs[i])
    axs[i].set_title('Histogram overlevend - {:0.1%} Survived ({})'.format(p[1], sex))
    axs[i].set_ylabel('% of passagiers')
    axs[i].set_xlabel('0 = dood, 1 = overleefd')
```