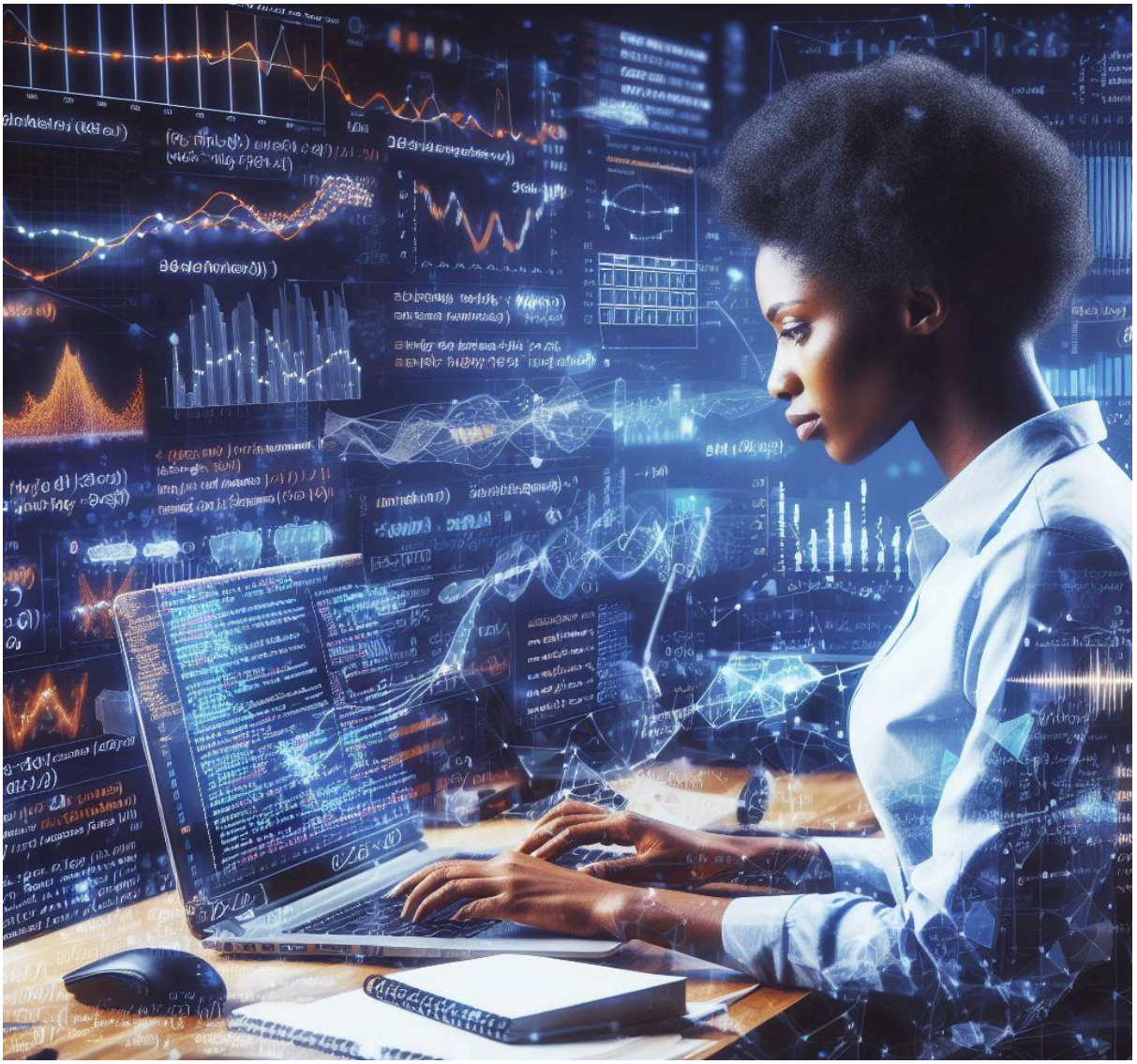
Dato: 26.10.2023
Versjon 1.0

NTNU
Kunnskap for en bedre verden

Analyse av datasett med multippel lineær regresjon



Microsoft Bing prompt: “Multiple Linear Regression” (Microsoft, 2023)

NTNU
Kunnskap for en bedre verden

Innholdfortegnelse

Sammendrag

Denne rapporten presenterer resultatene fra et prosjektarbeid i faget ISTT1003, statistikk ved NTNU, hvor formålet var å undersøke om LEGO-sett for gutter er dyrere enn for jenter. Prosjektet benytter multippel lineær regresjon på et datasett med 1304 observasjoner av individuelle LEGO-sett.

Analysen inkluderer en hypotesetest og evaluering av regresjonsmodellen ved bruk av bestemmelseskoeffisienten R^2, justert-R^2, residualplott og QQ-plott. Selv om modellen har høy bestemmelseskoeffisient ($R^2 = 0.852$), viser residual- og QQ-plottet begrensninger. Konklusjonen er at datasettet ikke gir tilstrekkelig grunnlag, og ytterligere tilpasninger, renseprosesser eller et større datasett er nødvendig for mer pålitelige resultater.

Introduksjon

Denne rapporten er skrevet på bakgrunn av et prosjektarbeid i faget ISTT1003 - Statistikk, ved institutt for datateknologi og informatikk, ved Norges tekniske naturvitenskapelige universitet. Prosjektarbeidet omhandler en analyse av et datasett ved bruk av multippel lineær regresjon. Datasettet, som er hentet fra Peterson og Ziegler (2021), inneholder 1304 observasjoner, hvor hver enkelt observasjon beskriver et individuelt LEGO-sett.

Formålet med denne rapporten er å undersøke om datasettet er tilstrekkelig nok til å kunne besvare den definerte problemstillingen: Er LEGO-sett for gutter dyrere enn for jenter? Gruppen vil gjennom grundige analyser av datasettet, hypotesetesting, evaluering og tilpasning av regresjonsmodellen avgjøre om de har nok grunnlag til å besvare problemstillingen.

Teori

Enkel lineær regresjon er en statistisk metode som beskriver sammenhengen mellom en respons og dens forklaringsvariabel. Multippel lineær regresjon utvider dette konseptet ved å introdusere flere forklaringsvariabler (Langaas, u.å). Dette innebærer at responsen nå er avhengig av et større og mer variert sett med data, som gir en mer omfattende modell og et bedre grunnlag for å estimere den.

I multippel lineær regresjon anvendes hypotesetesting for å teste om forklaringsvariablene har en signifikant effekt på responsen (Langaas, u.å). Man skiller mellom tosidig og ensidig hypotesetesting. En tosidig hypotesetest undersøker om det er en signifikant forskjell mellom to variabler eller om effekten er lik null, mens en ensidig hypotesetest brukes når man ønsker å undersøke om noe er større eller mindre enn noe annet. En hypotesetest gjennomføres ved å formulere en nullhypotese og en alternativ hypotese. Nullhypotesen er utgangspunktet og er hypotesen man ønsker å undersøke om man har grunnlag til å forkaste. Den alternative hypotesen er påstanden man ønsker å teste. Merk at hvis man ikke har grunnlag til å forkaste nullhypotesen, betyr det ikke nødvendigvis at man kan anta at nullhypotesen er sann. Det antyder heller at man ikke har tilstrekkelig grunnlag til å komme med en bestemt konklusjon basert på dataene (Bjørnland, u.å, "Uke 8").

Bestemmelseskoeffisienten, R^2, brukes ofte som en indikator i multippel lineær regresjon for å avgjøre hvor presis en regresjonsmodell er. R^2-verdien representerer andelen varians i responsen som modellen forklarer (Bjørnland, u.å, "Uke 9"). Desto nærmere verdien er 1, desto bedre passer modellen dataene. Merk at bestemmelseskoeffisienten ikke alene kan gi fullstendig informasjon om modellens kvalitet. Imidlertid, i kombinasjon med et residualplott og QQ-plott, kan den gi en bedre tilnærming om hvor godt modellen passer dataene (Sørmoen, 2023).

For å sammenligne to regresjonsmodeller anvendes justert-R^2, et mål som forklarer hvor godt modellen passer til populasjonen som datasettet stammer fra (Løvås, 2013). Den mest optimale modellen vil være den som oppnår høyest mulig R^2-verdi ved hjelp av færrest forklaringsvariabler.

Metode

For å svare på problemstillingen er det nødvendig å preprossesere datasettet, slik at det blir egnet for videre analyse. Første steg innebærer fjerning av kolonner som ikke skal brukes i oppgaven, og fjerne rader som inneholder ugyldige verdier. Dataen blir videre kategorisert under de ulike kjønnskategoriene. Dette skjer ved å gå gjennom en rekke forhåndsdefinerte koblinger mellom kjønn og tema, og legge til kjønnet i den nye kolonnen, "Gender". For å kategorisere legosett som mangler tema, går koden gjennom lister med jente- og guttenavn

(Kantrowitz, 1991a) & (Kantrowitz, 1991b) som sjekker for likheter for hvert legosett. I tillegg til dette går koden også gjennom en rekke nøkkelord som assosieres med de forskjellige kjønnene (Figur 1). Legosett som ikke inngår under lista får verdien “Neutral” i "Gender"-kolonnen.

Kategorisering på tema		
Gutt	**Jente**	**Nøytralt**
Vold	Prinsesser	Generell utdanning
Mekanikk (bil, båt, truck, tractor etc.)	Glitter, blomster, mye rosa/lilla	Ikke tema/generelle tema (arkitektur, brann, politi, ambulanse)
Datamaskin/konsollspill basert (minecraft, overwatch)		Ekstradeler
Superhelt er gutt	Superhelt er jente	
Alle/fleste karakterer er gutter	Alle/fleste karakterer er jenter	Alle/fleste karakterer er uten kjønn eller kjønnsnøytrale
Hovedkarakter er gutt	Hovedkarakter er jente	Hovedkarakter er uten kjønn eller kjønnsnøytrale

Kategorisering på nøkkelord			
Gutt	**Eks**	**Jente**	**Eks**
Navn	Kantrowitz datasett med 3000+ navn	Navn	Kantrowitz datasett med 3000+ navn
Mekanikk	Bil, båt, tractor	Prinsesse	Elsa, princess, elf
Gutt som hovedkarakter/er	Ninjago, avengers	Jente som hovedkarakter/er	Mermaid
Vold	fighter	Annet	Flower, bracelet, besties
Mannlig superhelt	Spiderman, batman		

Figur 1: Tabellene viser betraktninger som er gjort for å klassifisere datasettene.

For å undersøke sammenhengen mellom prisen på LEGO-sett og kjønn, ble det tatt i bruk en multippel lineær regresjonsmodell. I samsvar med problemstillingen blir prisen responsen i modellen, og de resterende variablene blir mulige forklaringsvariabler. For å kunne besvare problemstillingen ble det utledet en ensidig hypotesetest basert på regresjonsmodellen, der nullhypotesen antar at prisen for LEGO-sett for gutter er lik prisen på LEGO-sett for jenter, mens den alternative hypotesen er at prisen for LEGO-sett for gutter er dyrere enn for jenter.

Ved å kombinere ulike forklaringsvariabler og sammenligne med den justerte R^2-verdien, ble variablene “pieces”, “pages” og “gender” den kombinasjonen av variabler som ga høyest justert-R^2.

Resultater

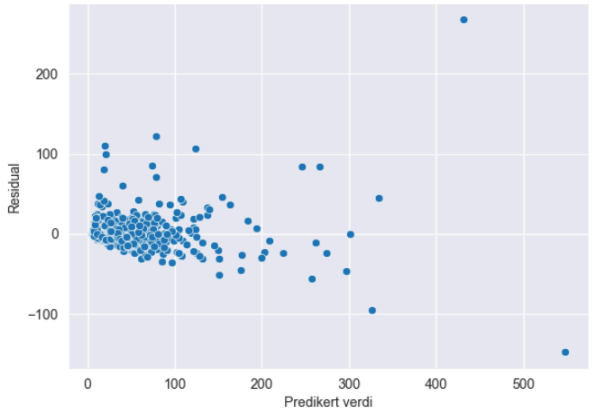
OLS Regression Results

Dep. Variable:	Price	R-squared:	0.852
Model:	OLS	Adj. R-squared:	0.850
Method:	Least Squares	F-statistic:	712.3
Date:	Thu, 09 Nov 2023	Prob (F-statistic):	1.31e-304
Time:	18:14:18	Log-Likelihood:	-3328.0
No. Observations:	752	AIC:	6670.
Df Residuals:	745	BIC:	6702.
Df Model:	6		
Covariance Type:	nonrobust		

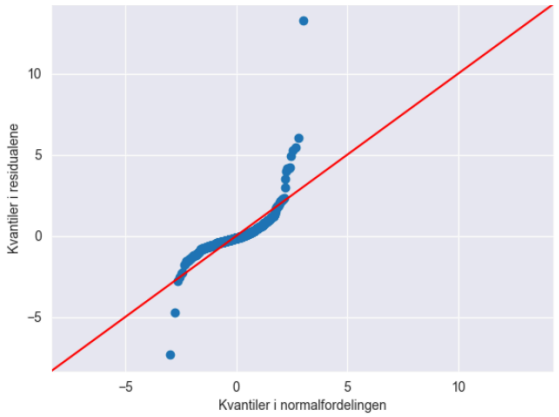
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.3389	3.054	0.766	0.444	-3.657	8.335
C(Gender, Treatment("Girl"))[T.Boy]	3.0206	3.327	0.908	0.364	-3.511	9.552
C(Gender, Treatment("Girl"))[T.Neutral]	5.9776	3.505	1.705	0.089	-0.903	12.858
Pages	0.0777	0.015	5.024	0.000	0.047	0.108
Pieces	0.0778	0.009	8.454	0.000	0.060	0.096
Pieces:C(Gender, Treatment("Girl"))[T.Boy]	0.0041	0.009	0.472	0.637	-0.013	0.021
Pieces:C(Gender, Treatment("Girl"))[T.Neutral]	-0.0069	0.009	-0.787	0.432	-0.024	0.010

Figur 2: Utskrift fra multippel lineær regresjon på observerte data i Python.

Fra tabellen ovenfor (Figur 2) kan vi lese av relevant informasjon. Dette inkluderer 752 ulike observasjoner, en betydelig høy R^2 - verdi på 0.852, samt t-verdier for de ulike forklaringsvariablene.

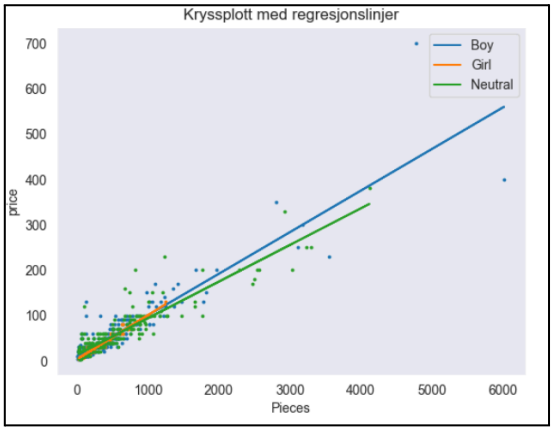


Figur 3: Utskriften beskriver et residualplott.

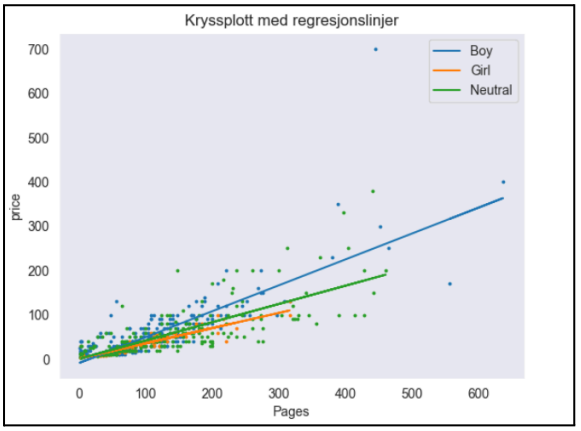


Figur 4: Utskriften til beskriver et QQ-plot.

Residualplottet (Figur 3) illustrerer en viss trend mellom predikert verdi og residual, ettersom punktene ikke er normalfordelt rundt 0-linjen, men heller er forskjøvet i negativ retning. I QQ-plottet (Figur 4) observerer man store avvik ved halene i fordelingen.



Figur 5: Pris i forhold til brikker.



Figur 6: Pris i forhold til sider.

Plottet (Figur 5) illustrerer sammenhengen mellom den kontinuerlige forklaringsvariablen, antall brikker, og responsen pris. Plottet (Figur 6) illustrerer en sammenheng mellom den kontinuerlige forklaringsvariabelen, sider, og pris.

Diskusjon

Fra den ensidige hypotesetesten nevnt i metodekapittelet, har vi ikke nok grunnlag for å forkaste nullhypotesen. Dette skyldes at testobservatoren, med en verdi på 0.472, ikke er større enn den kritiske verdien fra t-fordelingen. Dermed oppfylles ikke kravet $t > t_{\alpha, n-1}$ og testobservatoren befinner seg ikke i forkastningsområdet.

Resultatene fra modellene gir innsikt for å vurdere presisjonen til regresjonsmodellen. Bestemmelseskoeffisienten R^2, med de angitte forklaringsvariablene, forklarer rundt 85% av variasjonen i datasettet. Dette er en høy verdi, men er ikke tilstrekkelig nok for å vurdere presisjonen til modellen og må videre vurderes ved bruk av residual- og QQ-plottet.

Som nevnt i resultatet viser residualplottet at residualene ikke er normalfordelt rundt 0 linjen. I tillegg ser man at det finnes et fåtall legosett som har særdeles høye predikerte verdier, og som samtidig ikke virker normalfordelte. Dette tilsier at modellen kan være ganske svak for dyre legosett, og at modellen generelt ikke er helt normalfordelt. I QQ-plottet observerer man at halene er for tunge, noe som antyder at det finnes flere legosett med betydelig høy eller lav pris i forhold til det som var forventet.

Konklusjon

Basert på hypotesetesten, kan vi ikke konkludere om LEGO-sett for gutter er dyrere enn for jenter basert på datasettet. Imidlertid, ved å analysere residualplottet og QQ-plottet, kan vi i tillegg konkludere med at modellen ikke er optimal for å kunne besvare problemstillingen på en tilfredsstillende måte. Dermed er ikke datasettet tilstrekkelig for å kunne fastslå kvaliteten på konklusjonen med sikkerhet. Ytterligere undersøkelser og tilpasninger av modellen er nødvendig for et mer pålitelig resultat.

For å kunne styrke påliteligheten til modellen og komme til en mer sikker konklusjon, kunne man gjort ytterligere tilpasninger av modellen. Dette kunne f.eks. vært å rense dataene på en bedre eller en annen måte, finne alternative metoder for å kategorisere dataene på, gjøre nøyere undersøkelser på valg av forklaringsvariabler eller rett og slett bruke et større datasett for å få et mer representativt utvalg.

Literaturliste

Bjørnland, T. (u.å.) *Uke 9: 3 Inferens på stigningstallet i regresjonslinja, og hvor god er modellen?*. Tilgjengelig fra: https://ntnu.cloud.panopto.eu/Panopto/Pages/Viewer.aspx?id=af53d7c3-b389-4d31-8d7b-ad34007a1842 (Hentet: 30. Oktober 2023)

Bjørnland, T. (u.å.) *Uke 8: 1 Introduksjon til hypotesetesting*. Tilgjengelig fra: https://ntnu.cloud.panopto.eu/Panopto/Pages/Viewer.aspx?id=3416a1a8-b611-4965-a00b-ad340078c5b9 (Hentet: 30. Oktober 2023)

Kantrowitz, M. (1991a) *List of common female names.* Tilgjengelig fra: https://www.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/nlp/corpora/names/female.txt (Hentet: 6. November 2023).

Kantrowitz, M. (1991b) *List of common male names.* Tilgjengelig fra: https://www.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/nlp/corpora/names/male.txt (Hentet: 6. November 2023).

Langaas, M. (u.å.) *Multippel lineær regresjon: introduksjon.* Tilgjengelig fra: https://ntnu.cloud.panopto.eu/Panopto/Pages/Viewer.aspx?id=b2bb21a5-9a09-4ac7-aa25-ac5801055d5e (Hentet: 30. Oktober 2023)

Løvås, G.G. (2013) *Statistikk for universiteter og høgskoler.* 3. Utgave. Oslo: Universitetsforlaget

Microsoft. (2023) *DALL-E 3 (30. oktober-versjon)* [text-to-image model]. https://www.bing.com/create

Peterson, A. D. og Ziegler, L. (2021) *Building a Multiple Linear Regression Model With LEGO Brick Data*. Tilgjengelig fra: https://www.tandfonline.com/doi/full/10.1080/26939169.2021.1946450?scroll=top&needAccess=true (Hentet: 26. Oktober 2023).

Sørmoen, I. H. (2023) *Forelesning 2 ISTx1003.* Tilgjengelig fra: https://ntnu.cloud.panopto.eu/Panopto/Pages/Viewer.aspx?id=84e10f13-4cf6-4ef0-91f0-b0a500f58f58 (Hentet: 30. Oktober 2023)

Vedlegg

Utsnitt av kode:

```
import numpy as np

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

import statsmodels.formula.api as smf
import statsmodels.api as sm
import warnings
```

```
# Ignore SettingWithCopyWarning
warnings.filterwarnings('ignore')

df_2 = pd.read_csv("../resources/lego.population.csv", sep =",",
encoding ="latin1")

#Kiutter ut kolonner fra datasettet vi ikke trenger
df2 = df_2[['Set_Name', 'Theme', 'Pieces', 'Price', 'Pages',
'Unique_Pieces', 'Minifigures']]

# Gjør themes om til string og fjern alle tegn vi ikke vil ha med
df2['Theme'] = df2['Theme'].astype(str)
df2['Theme'] = df2['Theme'].str.replace(r'[^a-zA-Z0-9\s-]', '', regex
= True)

# Fjerner dollartegn og trademark-tegn fra datasettet
df2['Price'] = df2['Price'].str.replace('\$', '', regex = True)

# og gjør så prisen om til float
df2['Price'] = df2['Price'].astype(float)
```

```
#egen rens
df2['Gender'] = "Neutral"

#legger til boy names
boy_names = []

with open("../resources/boy_names.txt") as file:
  for index, line in enumerate(file):
    if index < 6:
      continue
    name = line.strip().lower()

    boy_names.append(name)

#legger til girl names
girl_names = []

with open("../resources/girl_names.txt") as file:
  for index, line in enumerate(file):
    if index < 6:
      continue
    name = line.strip().lower()

    girl_names.append(name)

#funksjon som sjekker om et navn finnes i listen over jenter
def inGirlList(set_name):
  for girl in girl_names:
    if str(set_name).find(girl) != -1:
      return "Girl"
  return "Neutral"

#funksjon som sjekker om et navn finnes i listen over gutter
def inBoyList(set_name):
  for boy in boy_names:
    if str(set_name).find(boy) != -1:
      return "Boy"
  return "Neutral"

#kategoriserer hvert stt etter om set_navnet finnes i jente- eller
guttelistene.
df2['Gender'] = df2['Set_Name'].apply(lambda x: inGirlList(x) if
inGirlList(x) == 1 else inBoyList(x))


#Legger til en rekke kobliner mellom tema og kjønn
themes = dict()
for line in open("../resources/theme_gender.txt"):
  if line.startswith("*") or line.isspace():
    continue
```

```
  else:
    words = line.strip().split("=")
    themes[words[0].lower().replace(" ", "")] = words[1].replace(" ",
"")

#setter nye gender-verdier for sett som har en theme
for index, lego_set in df_2.iterrows():
  set_theme = str(lego_set["Theme"]).lower().replace(" ", "")
  if set_theme in themes.keys():
    newValue = int((themes[set_theme]))

    if newValue == 1:
      df2['Gender'][index] = "Boy"


    elif newValue == 2:
      df2['Gender'][index] = "Girl"
    else:
      df2['Gender'][index] = "Neutral"

df2['Gender'] = pd.Categorical(df2['Gender'])

#dropper lego_sett som har NaN-verdier i en av kolonnene
df2 = df2.dropna()

df2
```

	Set_Name	Theme	Pieces	Price	Pages
13	Stephanie's Summer Heart Box	Friends	95.0	7.99	40.0
16	Woody & RC	Disney	69.0	9.99	28.0
17	Mia's Summer Heart Box	Friends	85.0	7.99	36.0
18	Olivia's Summer Heart Box	Friends	93.0	7.99	40.0
19	Police Patrol Car	City	92.0	9.99	36.0
...	...	...	...	...	...
1171	1989 Batmobile	Batman	3306.0	249.99	404.0
1172	Tree House	Ideas	3036.0	199.99	428.0
1173	Welcome to Apocalypseburg!	THE LEGO MOVIE 2	3178.0	299.99	452.0
1174	Jurassic Park: T. rex Rampage	Jurassic World	3120.0	249.99	464.0
1175	Monkie Kid's Team Secret HQ	Monkie Kid	1105.0	169.99	

```
556.0

      Unique_Pieces  Minifigures   Gender
13             52.0          1.0     Girl
16             36.0          1.0  Neutral
17             41.0          1.0     Girl
18             48.0          2.0     Girl
19             52.0          1.0  Neutral
...             ...          ...      ...
1171          484.0          3.0  Neutral
1172          482.0          4.0  Neutral
1173          692.0         13.0      Boy
1174          525.0          6.0      Boy
1175          622.0          7.0      Boy

[752 rows x 8 columns]
```

```
#multippel lineær modell
modell_mlr = smf.ols('Price ~ Pages + Pieces * C(Gender,
Treatment("Girl"))', data = df2)
print(modell_mlr.fit().summary())
```

```
                            OLS Regression Results

==============================================================================
========
Dep. Variable:                   Price   R-squared:
0.852
Model:                             OLS   Adj. R-squared:
0.850
Method:                  Least Squares   F-statistic:
712.3
Date:                 Thu, 16 Nov 2023   Prob (F-statistic):
1.31e-304
Time:                         16:59:55   Log-Likelihood:
-3328.0
No. Observations:                 752   AIC:
6670.
Df Residuals:                     745   BIC:
6702.
Df Model:                           6

Covariance Type:             nonrobust

==============================================================================
==============================================
                                                    coef    std err
t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
----------------------------------------------
```

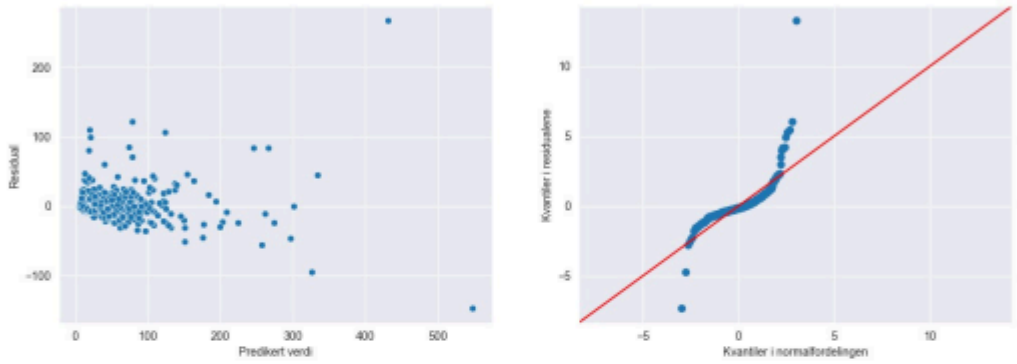
```
Intercept                                                 2.3389      3.054
0.766      0.444      -3.657       8.335
C(Gender, Treatment("Girl"))[T.Boy]                       3.0206      3.327
0.908      0.364      -3.511       9.552
C(Gender, Treatment("Girl"))[T.Neutral]                   5.9776      3.505
1.705      0.089      -0.903      12.858
Pages                                                     0.0777      0.015
5.024      0.000       0.047       0.108
Pieces                                                    0.0778      0.009
8.454      0.000       0.060       0.096
Pieces:C(Gender, Treatment("Girl"))[T.Boy]                0.0041      0.009
0.472      0.637      -0.013       0.021
Pieces:C(Gender, Treatment("Girl"))[T.Neutral]           -0.0069      0.009
-0.787      0.432      -0.024       0.010
==============================================================================
========
Omnibus:                        711.768   Durbin-Watson:
1.765
Prob(Omnibus):                    0.000   Jarque-Bera (JB):
80335.298
Skew:                             3.824   Prob(JB):
0.00
Kurtosis:                        53.054   Cond. No.
6.18e+03
==============================================================================
========

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[2] The condition number is large, 6.18e+03. This might indicate that
there are
strong multicollinearity or other numerical problems.
```

```
myGenders = ['Boy', 'Girl', 'Neutral']
subset_df = df2[df2['Gender'].isin(myGenders)]

#Enkel lineær regresjon for hvert kjønn hver for seg
resultater = []
for i, gender in enumerate(myGenders):
  sub_model_gender = smf.ols('Price ~ Pieces',
data=subset_df[subset_df['Gender'] == gender])
  resultater.append(sub_model_gender.fit())

# plott av dataene og regresjonslinjene
for i, gender in enumerate(myGenders):
  slope = resultater[i].params['Pieces']
  intercept = resultater[i].params['Intercept']

  regression_x = np.array(subset_df[subset_df['Gender'] == gender]
```

```
['Pieces'])
  regression_y = slope * regression_x + intercept

  # Plot scatter plot and regression line
  plt.scatter(subset_df[subset_df['Gender'] == gender]['Pieces'],
subset_df[subset_df['Gender'] == gender]['Price'],
color=plt.cm.tab10(i), s=3)
  plt.plot(regression_x, regression_y, color=plt.cm.tab10(i),
label=gender)

plt.xlabel('Antall brikker')
plt.ylabel('Pris $')


plt.title('Kryssplott med regresjonslinjer')
plt.legend()
plt.grid()


plt.show()
```

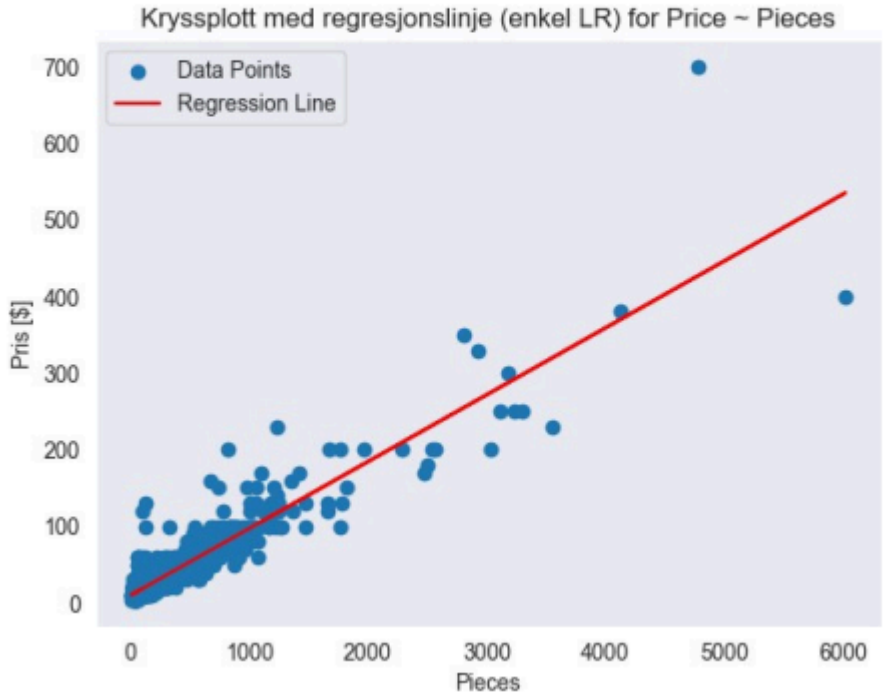

```
figure, axis = plt.subplots(1, 2, figsize = (15, 5))
sns.scatterplot(x = modell_mlr.fit().fittedvalues, y =
```

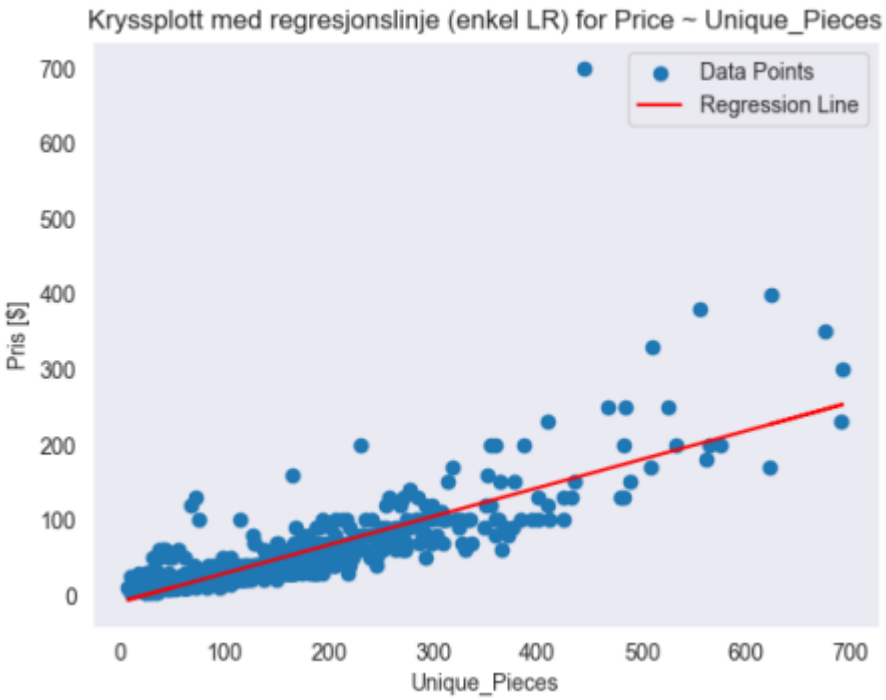
```
modell_mlr.fit().resid, ax = axis[0])
axis[0].set_ylabel("Residual")
axis[0].set_xlabel("Predikert verdi")

# Lage kvantil-kvantil-plott for residualene
sm.qqplot(modell_mlr.fit().resid, line = '45', fit = True, ax =
axis[1])
axis[1].set_ylabel("Kvantiler i residualene")
axis[1].set_xlabel("Kvantiler i normalfordelingen")
plt.show()
```

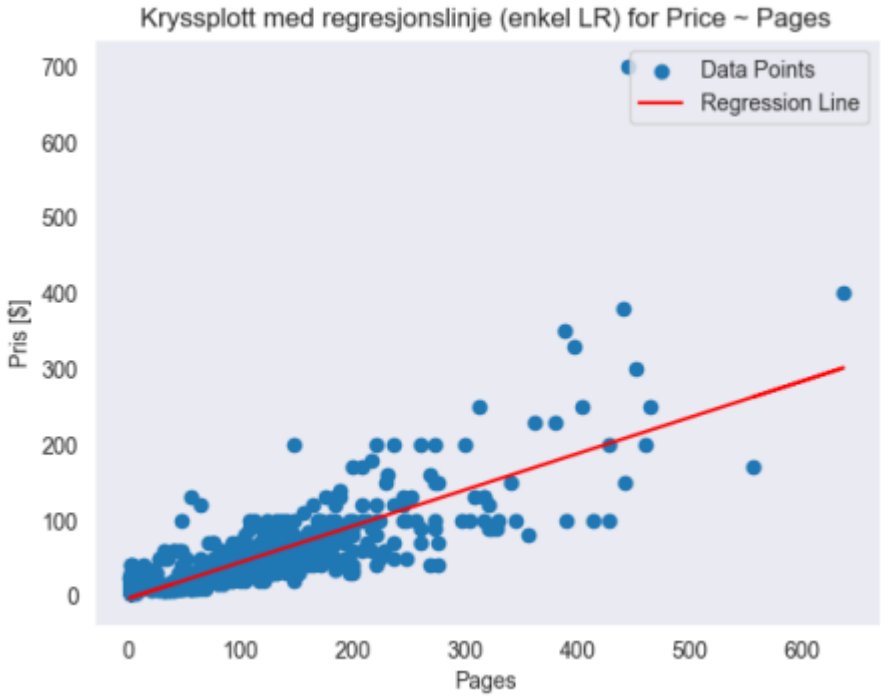


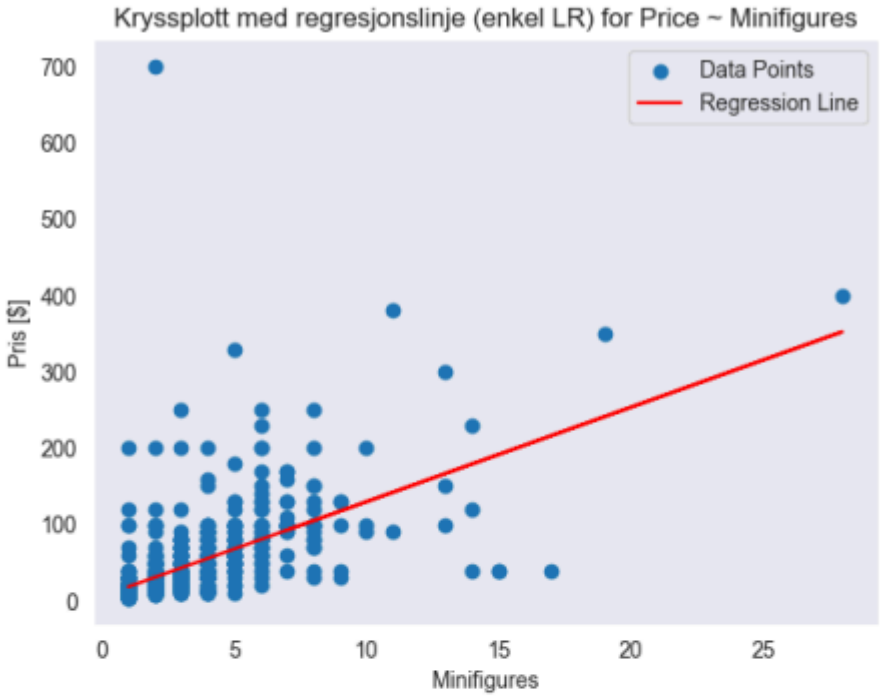

```
#Enkel lineær regresjon av alle de forskjellige variablene mot pris

variables = ['Pieces', 'Unique_Pieces', 'Pages', 'Minifigures']


for i in variables:
  formel = 'Price ~ ' + str(i)

  modell = smf.ols(formel, data = df2)
  resultat = modell.fit()

  resultat.summary()

  slope = resultat.params[i]
  intercept = resultat.params['Intercept']

  regression_x = np.array(df2[i])

  regression_y = slope * regression_x + intercept

  plt.scatter(df2[i], df2['Price'], label='Data Points')
  plt.plot(regression_x, regression_y, color='red', label='Regression
Line')
  plt.xlabel(i)
  plt.ylabel('Pris [$]')
```

```
plt.title('Kryssplott med regresjonslinje (enkel LR) for ' + formel)
plt.legend()
plt.grid()
plt.show()
```



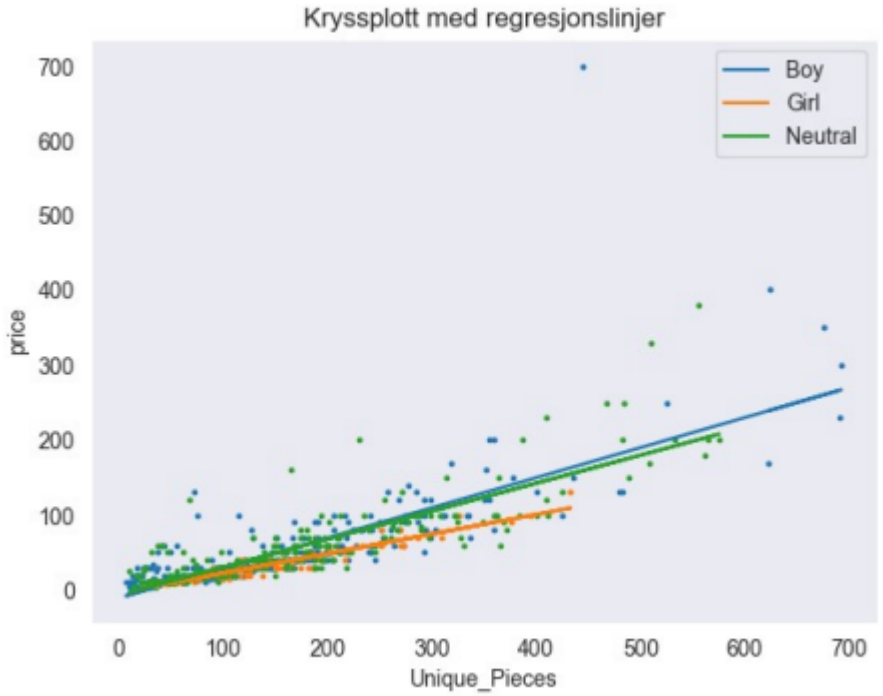


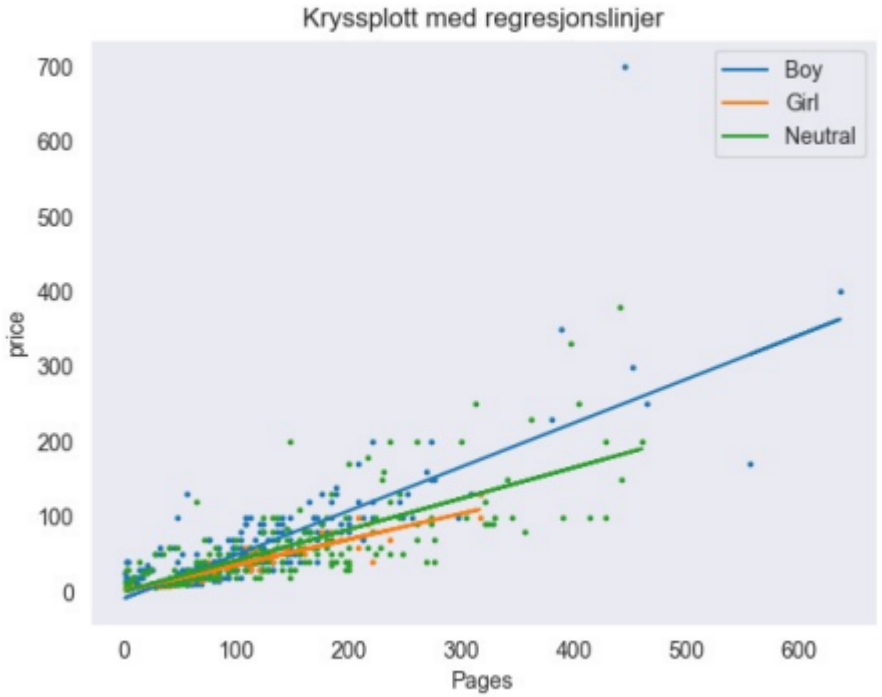

Kryssplott med regresjonslinje (enkel LR) for Price ~ Pages
Data Points
Regression Line
Pris [$]
Pages
0
100
200
300
400
500
600
700

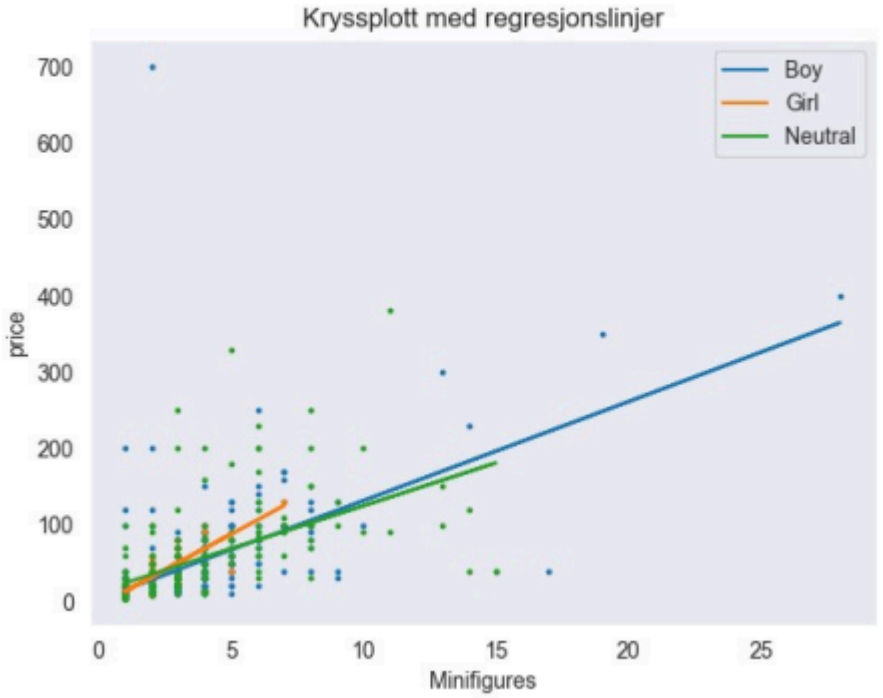



```
#Lineær regresjon for de forskjellige forklaringsvariablene, med en
regresjonslinje for hvert kjønn
for type in ['Pieces', 'Unique_Pieces', 'Pages', 'Minifigures']:
  myGenders = ['Boy', 'Girl', 'Neutral']
  subset_df = df2[df2['Gender'].isin(myGenders)]

  resultater = []
  for i, gender in enumerate(myGenders):
    sub_model_gender = smf.ols('Price ~ ' + type,
data=subset_df[subset_df['Gender'] == gender])
    resultater.append(sub_model_gender.fit())

  # plott av dataene og regresjonslinjene
  for i, gender in enumerate(myGenders):
    slope = resultater[i].params[type]
    intercept = resultater[i].params['Intercept']

    regression_x = np.array(subset_df[subset_df['Gender'] == gender]
[type])
    regression_y = slope * regression_x + intercept

    # Plot scatter plot and regression line
```

```
    plt.scatter(subset_df[subset_df['Gender'] == gender][type],
subset_df[subset_df['Gender'] == gender]['Price'],
color=plt.cm.tab10(i), s=3)
    plt.plot(regression_x, regression_y, color=plt.cm.tab10(i),
label=gender)

  plt.xlabel(type)
  plt.ylabel("price")


  plt.title('Kryssplott med regresjonslinjer')
  plt.legend()
  plt.grid()


  plt.show()
```


Kryssplott med regresjonslinjer
Boy
Girl
Neutral
price
Unique_Pieces
0
100
200
300
400
500
600
700


Kryssplott med regresjonslinjer
Boy
Girl
Neutral
price
Pages
0
100
200
300
400
500
600
700


Kryssplott med regresjonslinjer
Boy
Girl
Neutral
price
Minifigures
0
100
200
300
400
500
600
700
0
5
10
15
20
25