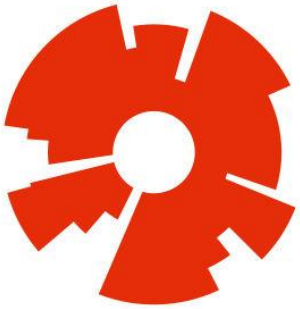


HÁSKÓLINN Í REYKJAVÍK
REYKJAVIK UNIVERSITY

I-707-VGBI

Viðskiptagreind

Hópverkefni 1

Teacher: Þórbergur Ólafsson

Arnar Már Kristinsson |arnark20@ru.is

Hinrik Pétur Jóhannsson | hinrik20@ru.is

Rut Tryggvadóttir |rutt20@ru.is

Inngangur

það er alltaf hluti af nemendum í öllu námi sem klára ekki námið. Það getur margt komið upp á hjá fólki, eins og barneignir, námsörðugleikar eða að fólk hreinlega missi áhugann á náminu. Öllu námi fylgir eitthvað brottfall, en spurningin sem við ætlum að reyna að svara í þessu verkefni er: “Hversu hátt hlutfall þeirra nemenda sem skrá sig í háskólanám klára ekki”.

Verkaskipting

Verkefnum var skipt jafnt og þétt á milli hópmeðlima og það var enginn einn sem vann meira en einhver annar. Arnar sá aðallega um Python, SQL vinnslu og skýrslugerð. Hinrik sá einnig um Python vinnslu og PowerPivot. Rut hafði yfirumsjón með því að sækja og hreinsa gögnin og sá um að þau væru rétt uppsett og væru að skrifast rétt inn í gagnagrunninn og PowerPivot.

Gögn

Gögnum var safnað um skráningu og útskrift hjá háskóla íslands frá árinu 2007 til ársins 2021, eða yfir fimmtán ára bil. Þessi gögn voru fengin af vef Háskóla Íslands (https://www.hi.is/kynningarefni/nemendur).
Þessi gögn urðu fyrir valinu vegna þess að okkur þóttu þau vera áhugaverð þar sem við erum öll í háskólanámi og líklegast höfum við öll hugsað á einhverjum tímapunkti hvort við ættum að pakka saman og hætta eftir slæman árangur í lokaprófi.
Út frá gögnunum sem var safnað er hægt að sjá hversu margar konur/karlar skrá sig í nám og útskrifast á hverri önn hjá Háskóla Íslands. Einnig er hægt að sjá hversu margir skrá sig eða útskrifast á hverju stigi í náminu, (grunnnám, viðbótarnám, meistarapróf, doktorspróf). Út frá þessum gögnum er hægt að segja margar skemmtilegar sögur, til dæmis hver hlutföllin eru á milli karla og kvenna í hverju námi fyrir sig, hvernig þróun hefur verið í skráningu í doktorsnám yfir árin eða hvernig vinsældir á námsbrautum hafa breyst undanfarin fimmtán ár. Við kusum að skoða tengslin á milli fjölda nemenda sem skrá sig í háskólanám og þeirra nemenda sem útskrifast og kanna hversu margir klára ekki háskólanámið, hvort sem það er grunnnám, viðbótarnám eða aðrar gráður.

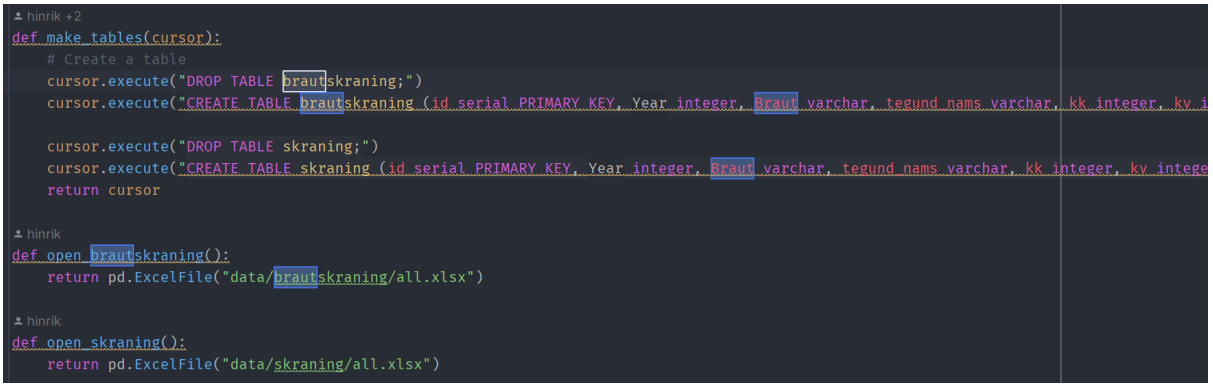
Framkvæmd

Vinnsla gagnanna gekk þannig fyrir sig að við byrjuðum á að hlaða niður öllum gögnunum fyrir skráningu og brautskráningu frá Háskóla Íslands á árunum 2006-2021. Næst settum við upp PostgreSQL gagnagrunn en PostgreSQL varð fyrir valinu vegna þess að það var sá gagnagrunnur sem hópmeðlimir höfðu mesta reynslu við að nota. Því næst hreinsuðum við gögnin lítillega til handvirkt til þess að hægt væri að nota ETL til að færa gögnin yfir í SQL/PowerPivot. Eftir að gögnin voru komin á staðlað form skrifuðum við Python skriftu sem hreinsaði gögnin enn meira og tengdist PostgreSQL gagnagrunninum og færði gögnin yfir í gagnagrunninn. Þegar gögnin voru komin inn í gagnagrunninn á stöðluðu formi gátum við byrjað að vinna gögnin. Við skrifuðum nokkrar einfaldar SQL skipanir til að fá það fram úr gögnunum sem myndi styðja við söguna okkar.

```python
# Connect to the database
connection = psycopg2.connect(user="postgres",
                              password="[redacted]",
                              host="localhost",
                              port="5433",
                              database="vidskiptagreind_hop1")


# # Create a dictionary cursor
cursor = connection.cursor()
# Print PostgreSQL details
print("PostgreSQL server information")
print(connection.get_dsn_parameters(), "\n")
# Executing a SQL query
cursor.execute("SELECT version();")
# Fetch result
record = cursor.fetchone()
print("You are connected to - ", record, "\n")
```

Python - Tenging við gagnagrunninn



```
hinrik +2
def make_tables(cursor):
    # Create a table
    cursor.execute("DROP TABLE brautskraning;")
    cursor.execute("CREATE TABLE brautskraning (id serial PRIMARY KEY, Year integer, Braut varchar, tegund_nams varchar, kk integer, kv i

    cursor.execute("DROP TABLE skraning;")
    cursor.execute("CREATE TABLE skraning (id serial PRIMARY KEY, Year integer, Braut varchar, tegund_nams varchar, kk integer, kv intege
    return cursor

hinrik
def open_brautskraning():
    return pd.ExcelFile("data/brautskraning/all.xlsx")

hinrik
def open_skraning():
    return pd.ExcelFile("data/skraning/all.xlsx")
```

Python - hreinsun gagna og SQL vinnsla 1

```
hinrik +2
def add_to_table(table, cursor, isbraut):
    sheet_list = table.sheet_names
    df_csv = {"Year": [], "Braut": [], "tegund_nams": [], "kk": [], "kv": [], "samtals": []}
    for sheet in sheet_list:
        print(sheet)
        df = table.parse(sheet)
        # print(df)
        df = df.dropna(thresh=4)
        df["Karl"].fillna(0, inplace=True)
        df["Kona"].fillna(0, inplace=True)
        df["Alls"].fillna(0, inplace=True)
        df = df.drop(columns=[df.columns[0]])
        # print(df)
        df = df.rename(columns={"Unnamed: 2": "tegund_nams"})

        df_isnan = df.isnull()

        temp_deild = ""
```

Python - hreinsun gagna og SQL vinnsla 2

```
for i in df["Karl"].keys():                                                    ▲ 2 ▲ 30 ✓ 49
    if df_isnan['Deild'][i] == False:
        temp_deild = df["Deild"][i]
        df_csv["Year"].append(sheet)
        df_csv["Braut"].append(df["Deild"][i])
        df_csv["tegund_nams"].append(None)
        df_csv["kk"].append(df["Karl"][i])
        df_csv["kv"].append(df["Kona"][i])
        df_csv["samtals"].append(df["Alls"][i])

        if isbraut:
            cursor.execute("INSERT INTO brautskraning (Year, Braut, tegund_nams, kk, kv, samtals) VALUES (%s, %s, %s, %s, %s, %s)"
                    (sheet, df["Deild"][i], None, df["Karl"][i], df["Kona"][i], df["Alls"][i]))
        else:
            cursor.execute("INSERT INTO skraning (Year, Braut, tegund_nams, kk, kv, samtals) VALUES (%s, %s, %s, %s, %s, %s)",
                    (sheet, df["Deild"][i], None, df["Karl"][i], df["Kona"][i], df["Alls"][i]))
    else:
        df_csv["Year"].append(sheet)
        df_csv["Braut"].append(temp_deild)
        df_csv["tegund_nams"].append(df["tegund_nams"][i])
        df_csv["kk"].append(df["Karl"][i])
        df_csv["kv"].append(df["Kona"][i])
        df_csv["samtals"].append(df["Alls"][i])
```

Python - hreinsun gagna og SQL vinnsla 3

```
            if isbraut:
                cursor.execute("INSERT INTO brautskraning (Year, Braut, tegund_nams, kk, kv, samtals) VALUES (%s, %s, %s, %s, %s, %s)",
                        (sheet, temp_deild, df["tegund_nams"][i], df["Karl"][i], df["Kona"][i], df["Alls"][i]))
            else:
                cursor.execute("INSERT INTO skraning (Year, Braut, tegund_nams, kk, kv, samtals) VALUES (%s, %s, %s, %s, %s, %s)",
                        (sheet, temp_deild, df["tegund_nams"][i], df["Karl"][i], df["Kona"][i], df["Alls"][i]))


    df_csv = pd.DataFrame(data=df_csv)
    if isbraut:
        df_csv.to_csv("braut_skra.csv", encoding="utf-8-sig")
    else:
        df_csv.to_csv("skraning_skra.csv", encoding="utf-8-sig")
    print("here")
    return cursor
```

Python - hreinsun gagna og SQL vinnsla 4

```
if __name__ == '__main__':
    cursor, connection = connect_to_postgres()
    braut = open_brautskraning()
    skraning = open_skraning()
    cursor = make_tables(cursor)
    cursor = add_to_table(braut, cursor, True)
    cursor = add_to_table(skraning, cursor, False)
    connection.commit()
```

Python - hreinsun gagna og SQL vinnsla 5

Þegar SQL vinnu var lokið fórum við yfir í næsta verkhluta sem var PowerPivot. Við skrifuðum aðra Python skriftu sem færði gögnin okkar úr .xls yfir í .csv skráarformat til þess að einfalda gagnavinnsluna í PowerPivot. Við vorum þegar búin að hreinsa gögnin svo vel til í python að okkur fannst ekki þörf á að hreinsa þau neitt frekar þegar þau voru komin inn í PowerPivot. Við tengdum gögnin og útbjuggum þau gröf sem studdu við söguna sem við ætluðum að segja með gögnunum.

Í næsta kafla verður fjallað um niðurstöðurnar sem fengust út frá þessari gagnavinnslu.

Niðurstöður

SQL

Við gerðum mestmegnið af gagnagreiningunni okkar í SQL en þar skoðuðum við hlutföll á milli nemenda sem skráðu sig og nemenda sem útskrifuðust og brutum það niður í nokkra flokka: eftir árum, deildum og kynjum.

Hlutfall þeirra sem skráðu sig á móti þeim sem útskrifuðust:

SQL Skipun:

```
SELECT
 SUM(s.skradir) as samtals_skradir,
 SUM(b.brautskradir) as samtals_brautskradir,
 ROUND((SUM(b.brautskradir) * 100.0 / SUM(s.skradir)), 2)
as utskriftarprosenta
FROM
 (SELECT SUM(samtals) as skradir
  FROM skraning
  WHERE tegund_nams is null) s,
 (SELECT SUM(samtals) as brautskradir
  FROM brautskraning
  WHERE tegund_nams is null) b;
```

Niðurstaða:

	samtals_skradir	samtals_brautskradir	utskriftarprosenta
1	181320	40739	22.47

Hlutfall þeirra sem skráðu sig á móti þeim sem útskrifuðust, skipt upp eftir árum:

SQL Skipun:

```
SELECT s.Year, s.skradir, b.brautskradir,
ROUND((b.brautskradir * 100.0 / s.skradir)::numeric, 2) as
utskriftarprosenta
FROM
 (SELECT Year, SUM(samtals) as skradir
  FROM skraning
  WHERE tegund_nams IS NULL
  GROUP BY Year) s
LEFT JOIN
 (SELECT Year, SUM(samtals) as brautskradir
  FROM brautskraning
  WHERE tegund_nams IS NULL
  GROUP BY Year) b
ON s.Year = b.Year
ORDER BY s.Year;
```

Niðurstöður:

	year	skradir	brautskradir	utskriftarprosenta
1	2007	9485	1795	18.92
2	2008	12048	1846	15.32
3	2009	13569	2261	16.66
4	2010	10068	2676	26.58
5	2011	9981	2780	27.85
6	2012	9929	2732	27.52
7	2013	10433	2717	26.04
8	2014	10595	2893	27.31
9	2015	13048	3172	24.31
10	2016	13107	3018	23.03
11	2017	13348	2956	22.15
12	2018	12901	2702	20.94
13	2019	12926	2949	22.81
14	2020	13835	2871	20.75
15	2021	16047	3371	21.01

Hlutfall þeirra sem skráðu sig á móti þeim sem útskrifuðust, skipt upp eftir brautum:

SQL Skipun:

```
SELECT s.braut, s.skradir, b.brautskradir,
ROUND((b.brautskradir * 100.0 / s.skradir)::numeric, 2)
as utskriftarprosenta
FROM
 (SELECT braut, SUM(samtals) as skradir
  FROM skraning
  WHERE tegund_nams IS NULL
  GROUP BY braut) s
LEFT JOIN
   (SELECT braut, SUM(samtals) as brautskradir
    FROM brautskraning
    WHERE tegund_nams IS NULL
    GROUP BY braut) b
ON s.braut = b.braut
ORDER BY s.braut;
```

Niðurstöður:

	braut	skradir	brautskradir	utskriftarprosenta
1	Deild erlendra tungumála, bókmennta og má…	10010	1648	16.46
2	Félags- og mannvísindadeild	13551	3021	22.29
3	Félagsráðgjafardeild	5506	1547	28.1
4	Guðfræði- og trúarbragðafræðideild	1448	428	29.56
5	Hagfræðideild	4549	693	15.23
6	Hjúkrunarfræðideild	7191	1910	26.56
7	Iðnaðarverkfræði-, vélaverkfræði- og tölv…	9704	2379	24.52
8	Íslensku- og menningardeild	13484	2687	19.93
9	Íþrótta-, tómstunda- og þroskaþjálfadeild	6317	1523	24.11
10	Jarðvísindadeild	2564	501	19.54
11	Kennaradeild	17473	4010	22.95
12	Lagadeild	8209	2162	26.34
13	Líf- og umhverfisvísindadeild	6087	1353	22.23
14	Lyfjafræðideild	2304	744	32.29
15	Læknadeild	9958	2580	25.91
16	Matvæla- og næringarfræðideild	1630	330	20.25
17	Rafmagns- og tölvuverkfræðideild	1594	355	22.27
18	Raunvísindadeild	4328	922	21.3
19	Sagnfræði- og heimspekideild	7665	1153	15.04
20	Sálfræðideild	7866	1765	22.44
21	Stjórnmálafræðideild	9661	2092	21.65
22	Tannlæknadeild	1111	157	14.13
23	Umhverfis- og byggingaverkfræðideild	2382	506	21.24
24	Uppeldis- og menntunarfræðideild	8345	1687	20.22
25	Viðskiptafræðideild	18383	3992	21.72

Hlutfall þeirra sem skráðu sig á móti þeim sem útskrifuðust, skipt upp eftir kynjum:

SQL Skipun:

```
SELECT s.Year, s.kk_skradir, s.kvk_skradir,
b.kk_brautskradir, b.kvk_brautskradir,
ROUND((b.kk_brautskradir * 100.0 /
s.kk_skradir)::numeric, 2) as utskriftarprosenta_kk,
ROUND((b.kvk_brautskradir * 100.0 /
s.kvk_skradir)::numeric, 2) as prosenta_kvk
FROM
 (SELECT Year, SUM(kk) as kk_skradir, SUM(kv) as
kvk_skradir
  FROM skraning
  WHERE tegund_nams IS NULL
  GROUP BY Year) s
LEFT JOIN
 (SELECT Year, SUM(kk) as kk_brautskradir, SUM(kv) as
kvk_brautskradir
  FROM brautskraning
```

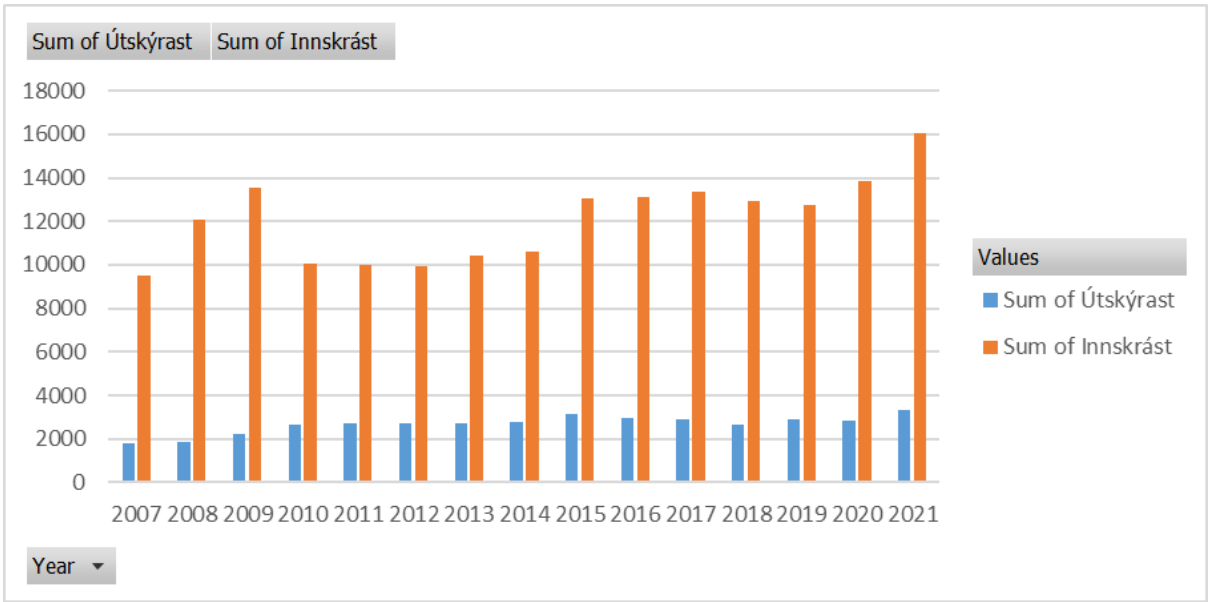
```
  WHERE tegund_nams IS NULL
  GROUP BY Year) b
ON s.Year = b.Year
ORDER BY s.Year;
```

Niðurstöður:

	year	kk_skradir	kvk_skradir	kk_brautskradir	kvk_brautskradir	utskriftarprosenta_kk	prosenta_kvk
1	2007	2938	6545	610	1205	20.76	18.41
2	2008	3894	8152	609	1248	15.64	15.31
3	2009	4391	9178	647	1630	14.73	17.76
4	2010	3289	6760	769	1974	23.38	29.2
5	2011	3354	6627	857	1941	25.55	29.29
6	2012	3341	6588	873	1859	26.13	28.22
7	2013	3574	6859	854	1863	23.89	27.16
8	2014	3566	6926	969	1948	27.17	28.13
9	2015	4496	8552	1100	2077	24.47	24.29
10	2016	4355	8752	1013	2009	23.26	22.95
11	2017	4535	8813	921	2052	20.31	23.28
12	2018	4364	8529	835	1881	19.13	22.05
13	2019	4410	8514	932	2031	21.13	23.85
14	2020	4488	9347	920	1963	20.5	21
15	2021	5184	10863	994	2383	19.17	21.94

Excel power pivot

Við byrjuðum að exporta gögnunum úr .xls yfir í .csv skráarformat til að geta unnið með þau í Power Pivot. Við náðum að tengja töflurnar inn í PowerPivot og gátum bætt þeim við sem gagnasettum en svo reyndum við að sameina töflurnar okkar í eitt rit en tókst það ekki. Við enduðum á að afrita dálkana sem við vildum í nýja töflu til að fá rétt rit. Hér fyrir neðan er tafla sem sýnir nemendur sem voru að skrá sig í nám (appelsínugulir dálkar) og þá sem voru að útskrifast (bláir dálkar) flokkað eftir árum.



PowerPivot - Rit sem sýnir útskriftir og skráningar flokkað eftir árum.

Samantekt á niðurstöðum

Okkur þóttu þessar niðurstöður mjög áhugaverðar og tölurnar voru mun lægri en við vorum búin að gera ráð fyrir. Það voru aðeins 22,4% af þeim sem skrá sig í háskóla sem klára. Árin 2007, 2008 og 2009 voru aðeins 16,9% nemenda sem útskrifuðust. En á næstu þremur árum eftir það, á árunum 2010, 2011 og 2012, var meðaltalið komið upp í 27,31%.

Ef við skoðum niðurstöðurnar eftir deildum er sú deild sem útskrifar flesta hlutfallslega Lyfjafræðideild þar sem hafa að meðaltali útskrifast 32.9%. Sú deild sem útskrifar fæsta hlutfallslega er Hagfræðideild sem útskrifar ekki nema 15,23% nemenda að meðaltali.

Ef við skoðum hlutföllin á milli karla og kvenna þá er mikill munur á milli ára, til dæmis voru aðeins 14,73% karla sem útskrifuðust árið 2009 á meðan það var 17,76% kvenna sem útskrifuðust sama ár. Svo árið 2014 var afskaplega gott útskriftarár þar sem 27,17% karla útskrifuðust og 28,13% kvenna.

Ef við hefðum haft meiri tíma hefði verið skemmtilegt að setja gögnin inn í kraftmeiri greiningartól til að greina betur mynstur og stefnur í gögnunum.

Álit hóps á verkefni

Okkur hópnum þótti þetta skemmtilegt og fræðandi verkefni og okkur þótti þetta mjög gott verkefni til að sýna hversu öflug þessi greiningartól geta verið og kostina og gallana á milli þeirra. Við lærðum helling á SQL, Excel, PowerPivot, CSV, Python og almenna gagnavinnslu með þessum tólum. Við vorum svolítið lengi að átta okkur á því hvað það var sem þurfti að gera og fór svolítill tími í að kynnast tólunum og setja upp umhverfið. En þegar allt var komið í gang gekk þetta almennt smurt fyrir sig.

Því miður náðum við ekki að finna nógu vel út úr PowerPivot og náðum við ekki að gera allt sem við ætluðum okkur í því. Þess vegna var langmest af gagnavinnslunni í Python/SQL. Hlutur sem tók okkur langan tíma að finna út úr hvernig við áttum að gera í SQL tók ekki nema nokkur handtök að græja í PowerPivot. Það er greinilegt að þetta er gríðarlega öflugt greiningartól sem við munum pottþétt nýta okkur í framtíðinni.

Lokaorð

Fyrir þetta verkefni var byrjað á því að velja gögnin sem við ætluðum að nota. Ákveðið var að nota gögn um skráningu og útskrift frá Háskóla Íslands. Niðurstöðurnar sem við vorum að reyna fá far hlutfall útskrifaðra miðað við skráningu í Háskóla Íslands. Hlutfallið á milli þessa tveggja liða endaði á að vera 22.4%. Þetta þýðir að það er stór hluti af nemendum sem skrá sig úr skóla áður en þau ná að klára. Það geta verið margar ástæður fyrir þessu, barneignir eða hreinlega áhugasviðsbreytingar. En brottfall úr skóla hefur almennt séð minnkað með árunum og við fögnum því.

Heimildaskrá

Háskóli Íslands. (e.d.). Nemendur - Kynningarefni.
https://www.hi.is/kynningarefni/nemendur

PyNative. (2021, 9. mars). Python Select from PostgreSQL Table using Psycopg2.
https://pynative.com/python-postgresql-select-data-from-table/

Viðbót

Í skjölunum sem fylgja með á canvas setjum við þessa skýrslu, SQL fyrirspurnirnar okkar í .sql skrá, Excel skjal sem inniheldur gögnin okkar í PowerPivot og myndbandskynningu á verkefninu. Við bætum einnig við Python skrá sem við notuðum til að auðga gagnagrunninn og ná gagnagrunnstengingunni.