COMPSS 214 1B Final Project

February 6, 2025

```
[ ]: import pandas as _hex_pandas
     import datetime as _hex_datetime
     import json as _hex_json
```

```
[ ]: hex_scheduled = _hex_json.loads("false")
```

```
[ ]: hex_user_email = _hex_json.loads("\"example-user@example.com\"")
```

```
[ ]: hex_user_attributes = _hex_json.loads("{}")
```

```
[ ]: hex_run_context = _hex_json.loads("\"logic\"")
```

```
[ ]: hex_timezone = _hex_json.loads("\"UTC\"")
```

```
[ ]: hex_project_id = _hex_json.loads("\"02b95162-ae32-470b-93e0-7a06bcacd212\"")
```

```
[ ]: hex_project_name = _hex_json.loads("\"COMPSS 214 1B: Final Project\"")
```

```
[ ]: hex_status = _hex_json.loads("\"\"")
```

```
[ ]: hex_categories = _hex_json.loads("[]")
```

```
[ ]: hex_color_palette = _hex_json.
     ↪loads("[\"#4C78A8\",\"#F58518\",\"#E45756\",\"#72B7B2\",\"#54A24B\",\"#EECA3B\",\"#B279A2\"
```

1 Impact of Family Structure & Education on the Gender Wage Gap in Management Roles

1.1 Research Question and Hypotheses

Background:

Research Question: What is the impact of family structure and education on the gender wage gap for individuals in management positions?

Measure: Yearly salary

Hypothesis 1: (Chi-Squared Test and Multiple Regression)

Null: Traditional family structures (partnered with children and married) do not significantly impact the gender wage gap and gender representation in management careers.

Alternative: Traditional family structures (partnered with children and married) have a significant impact on the gender wage gap or gender representation in management careers.

Variables of Interest: 'marind', 'gender', 'sector', 'salary', 'principal_job', 'leadership_job_titles', 'chlvin', 'ch6in', 'ch611in', 'ch1218in', 'ch19in', 'emsize', 'hrswk', 'agegr, 'newbus'

We are choosing to do a multiple regression and chi-squared test because the chi-squared test works with our categorical variables related to family structure and female representation to get the representation gap. We also use multiple regression to test the impacts on the gender wage gap.

Hypothesis 2: (Mediation Analysis)

Null: Education level does not mediate the gender wage gap in management careers for individuals in the business sector, holding marriage, children, and other factors constant.

Alternative: Education levels mediate the gender wage gap in management careers for individuals in the business sector, holding marriage, children, and other factors constant.

Variables of Interest: 'salary', 'gender', 'chlvin', 'marind', 'agegr', 'years_at_job', 'highest_degree_type', 'principal_job', 'leadership_job_titles', 'sector', 'emsize', 'hrswk', 'emrg', 'newbus', 'earn'

We are choosing a mediation analysis because education level can impact salary, since higher degrees may lead to higher wages in management roles, where often times, MBAs and PhDs are important. Additionally, education may mediate the gender wage gap if women's degrees are valued differently in the labor market. We wish to explore how much of the wage gap is explained by education and how much remains due to other factors.

1.1.1 Importing Packages and Data, Data Cleaning

[ ]:
```
import pandas as pd
import numpy as np
from matplotlib import pyplot
import seaborn as sns
import matplotlib.pyplot as plt
import csv
```

Read in the data. This code is the same as the code from class.

[ ]:
```
# import jinja2
# raw_query = """
#     select * from nscg_2021.csv
# """
# sql_query = jinja2.Template(raw_query).render(vars())
```

Read in codebook to get variables and their definitions/questions. This code is the same as the code from class.

[ ]:
```
codebook = pd.read_excel('Dpcg21.xlsx')
```

[ ]:
```
/home/hexuser/.cache/pypoetry/virtualenvs/python-kernel-
OtKFaj5M-py3.9/lib/python3.9/site-packages/openpyxl/styles/stylesheet.py:226:
UserWarning: Workbook contains no default style, apply openpyxl's default
  warn("Workbook contains no default style, apply openpyxl's default")
```

[ ]:
```
# import jinja2
# raw_query = """
#     select * from codebook
# """
# sql_query = jinja2.Template(raw_query).render(vars())
```

This is the same SQL code from class used to code variables for analysis.

[ ]:
```
# import jinja2
# raw_query = """
#     select
#         cast(salary as int) as salary, --capped at 9999996 - As of the week␣
 ↪of February 1, 2021, what was your basic annual salary on your principal␣
 ↪job, before deductions?
#         if (
#             cast(wkswk as int) = 98,
#             null,
#             cast(wkswk as int)
#         ) as wkswk, --Principal job salary: weeks per year basis; 98 is a␣
 ↪logical skip
#         if (
#             cast(earn as int) = 9999998,
#             null,
#             cast(earn as int)
#         ) as earn, --capped at 9999996 Total earned income before deductions␣
 ↪in previous year so need to also select those who worked last year
#         supwk,  --Supervised others in principal job indicator
#         wamgmt, --Work activity on principal job 10% indicator managing or␣
 ↪supervising people/project
#         case
#             when
#                 cast(supdir as int) = 9998
#                 then null
#             else cast(supdir as int)
#         end as supdir,
#         case
#             when
#                 cast(supind as int) = 99998
#                 then null
```

```
#             else cast(supind as int)
#         end as supind,
#         case
#             when cast(emsmi as int) = 1
#                 then 'same_employer_same_job'
#             when cast(emsmi as int) = 2
#                 then 'same_employer_different_job'
#             when cast(emsmi as int) = 3
#                 then 'different_employer_same_job'
#             when cast(emsmi as int) = 4
#                 then 'different_employer_different_job'
#             else null
#         end as emsmi,  --During these two time periods - the week of February 1, 2019, and the week of February 1, 2021 were you working for?
#         if(
#             gender = 'F',
#             1,
#             0
#         )
#         as gender,
#         white,
#         case
#             when cast(racethm as int) = 1
#                 then 'asian'
#             when cast(racethm as int) = 2
#                 then 'american_indian_alaskan'
#             when cast(racethm as int) = 3
#                 then 'black'
#             when cast(racethm as int) = 4
#                 then 'hispanic'
#             when cast(racethm as int) = 5
#                 then 'white'
#             when cast(racethm as int) = 6
#                 then 'pacific_islander'
#             when cast(racethm as int) = 7
#                 then 'multiple_race_non_hispanic'
#             end as racethm,
#         case
#             when resploc = '01'
#                 then 'new_england'
#             when resploc = '02'
#                 then 'middle_atlantic'
#             when resploc = '03'
#                 then 'east_north_central'
#             when resploc = '04'
#                 then 'west_north_central'
#             when resploc = '05'
```

```
#                 then 'south_atlantic'
#             when resploc = '06'
#                 then 'east_south_central'
#             when resploc = '07'
#                 then 'west_south_central'
#             when resploc = '08'
#                 then 'mountain'
#             when resploc = '09'
#                 then 'pacific_us_territories'
#             when resploc = '10'
#                 then 'europe'
#             when resploc = '20'
#                 then 'asia'
#             when resploc = '33'
#                 then 'caribbean'
#             when resploc = '37'
#                 then 'south_america'
#             when resploc = '40'
#                 then 'africa'
#             when resploc = '55'
#                 then 'abroad_unspecified'
#             end as resploc,
#         ctzusin, --us citizen
#         cast(refyr as int) as refyr, --year of survey
#         cast(biryr as int) as biryr, --birth year
#         agegr, --Age Group (5 year intervals)
#         cast(strtyr as int) as strtyr, --Year principal job started
#         prmbr, --Number of professional society memberships
#         promtgi, -- attended prof meetings
#         wktrni, -- attended trainings
#         if (cast(hrswk as int) = 98,
#             null,
#             hrswk) as hrswk, --capped at 96 hours a week
#         case
#             when cast(n2ocprbg as int) = 1
#                 then 'S&E'
#             when cast(n2ocprbg as int) = 2
#                 then 'S&E_related'
#             when cast(n2ocprbg as int) = 3
#                 then 'non_S&E'
#             else null
#             end as n2ocprbg, --Job code for principal job broad group SE based
#         case
#             when cast(n2ocprmg as int) = 1
#                 then 'computer_scientists'
#             when cast(n2ocprmg as int) = 2
#                 then 'bio_life_scientists'
```

```
#             when cast(n2ocprmg as int) = 3
#                 then 'physical_scientists_related'
#             when cast(n2ocprmg as int) = 4
#                 then 'social_scientists_related'
#             when cast(n2ocprmg as int) = 5
#                 then 'engineers'
#             when cast(n2ocprmg as int) = 6
#                 then 'S&E_related_occupations'
#             when cast(n2ocprmg as int) = 7
#                 then 'non-S&E_occupations'
#             else null
#             end as occupation_group,
#         n3ocpr, --Job code for principal job best code, includes Top-level␣
 ↪managers, execs, admins (711410) and OTHER mid-level managers (711470)
#         case when
#             n3ocpr in (621420, 621430,621440, 621450, 711460, 711470)
#                 then 'mid_level_manager'
#             when n3ocpr = 711410
#                 then 'c_suite_manager'
#             when n3ocpr = 721520
#                 then 'personnel_mgmt'
#             when n3ocpr = 721510
#                 then 'financial_mgmt'
#             when n3ocpr = 721530
#                 then 'other_management_occupations'
#             else null
#             end as leadership_job_codes,
#         case when
#             n3ocpr = 621420
#                 then 'computer_it_managers'
#             when n3ocpr = 621430
#                 then 'eng_managers'
#             when n3ocpr = 621440
#                 then 'medical_health_manager'
#             when n3ocpr = 621450
#                 then 'natural_science_manager'
#             when n3ocpr = 711460
#                 then 'ed_administrators'
#             when n3ocpr = 711470
#                 then 'other_mid_level_manager'
#             when n3ocpr = 711410
#                 then 'c_suite_manager'
#             when n3ocpr = 721520
#                 then 'personnel_mgmt'
#             when n3ocpr = 721510
#                 then 'financial_mgmt'
#             when n3ocpr = 721530
```

```
#                 then 'other_management_occupations'
#             else null
#             end as leadership_job_titles,
#         case
#             when cast(n3ocprng as int) = 11 then␣
 ↪'computer_information_scientists'
#             when cast(n3ocprng as int) = 12 then 'mathematical_scientists'
#             when cast(n3ocprng as int) = 18 then 'postsecondary␣
 ↪teachers_computer_and_math'
#             when cast(n3ocprng as int) = 21 then␣
 ↪'agricultural_food_scientists'
#             when cast(n3ocprng as int) = 22 then␣
 ↪'biological_medical_scientists'
#             when cast(n3ocprng as int) = 23 then␣
 ↪'environmental_life_scientists'
#             when cast(n3ocprng as int) = 28 then␣
 ↪'postsecondary_teachers_life_related_sciences'
#             when cast(n3ocprng as int) = 31 then 'chemists_except_biochemists'
#             when cast(n3ocprng as int) = 32 then␣
 ↪'earth_atmospheric_ocean_scientists'
#             when cast(n3ocprng as int) = 33 then 'physicists'
#             when cast(n3ocprng as int) = 34 then␣
 ↪'other_physical_related_scientists'
#             when cast(n3ocprng as int) = 38 then␣
 ↪'postsecondary_teachers_physical_and_related_sciences'
#             when cast(n3ocprng as int) = 41 then 'economists'
#             when cast(n3ocprng as int) = 42 then 'political_scientists'
#             when cast(n3ocprng as int) = 43 then 'psychologists'
#             when cast(n3ocprng as int) = 44 then 'anthropologists'
#             when cast(n3ocprng as int) = 45 then␣
 ↪'other_social_related_scientists'
#             when cast(n3ocprng as int) = 48 then␣
 ↪'postsecondary_teachers_social_related_sciences'
#             when cast(n3ocprng as int) = 51 then␣
 ↪'aerospace_aeronautical_astronautical_engineers'
#             when cast(n3ocprng as int) = 52 then 'chemical_engineers'
#             when cast(n3ocprng as int) = 53 then␣
 ↪'civil_architectural_sanitary_engineers'
#             when cast(n3ocprng as int) = 54 then␣
 ↪'electrical_computer_hardware_engineers'
#             when cast(n3ocprng as int) = 55 then 'industrial_engineers'
#             when cast(n3ocprng as int) = 56 then 'mechanical_engineers'
#             when cast(n3ocprng as int) = 57 then 'other_engineers'
#             when cast(n3ocprng as int) = 58 then␣
 ↪'postsecondary_teachers_engineering'
#             when cast(n3ocprng as int) = 61 then 'health_related_occupations'
```

```
#             when cast(n3ocprng as int) = 62 then 'S&E_managers'
#             when cast(n3ocprng as int) = 63 then 'S&E_pre_college_teachers'
#             when cast(n3ocprng as int) = 64 then␣
 ↪'S&E_technicians_and_technologists'
#             when cast(n3ocprng as int) = 65 then␣
 ↪'other_S&E_related_occupations'
#             when cast(n3ocprng as int) = 71 then 'non_S&E_managers'
#             when cast(n3ocprng as int) = 72 then␣
 ↪'management_related_occupations'
#             when cast(n3ocprng as int) = 73 then 'non_S&E_precollege_teachers'
#             when cast(n3ocprng as int) = 74 then␣
 ↪'non_S&E_postsecondary_teachers'
#             when cast(n3ocprng as int) = 75 then␣
 ↪'social_services_related_occupations'
#             when cast(n3ocprng as int) = 76 then␣
 ↪'sales_marketing_occupations'
#             when cast(n3ocprng as int) = 77 then␣
 ↪'art_humanities_related_occupations'
#             when cast(n3ocprng as int) = 78 then 'other_non_S&E_occupations'
#             else null
#             end as principal_job,
#         case
#             when wapri = '01'
#             then 'accounting_finance_contracts'
#             when wapri = '02'
#             then 'basic_research'
#             when wapri = '03'
#             then 'applied_research'
#             when wapri = '04'
#             then 'reasearch_dev_materials_devices'
#             when wapri = '05'
#             then 'design_equipment_processes_structures_models'
#             when wapri = '06'
#             then 'computer_apps_programming_systems_dev'
#             when wapri = '07'
#             then 'human_resources'
#             when wapri = '08'
#             then 'managing_supervising_people_projects'
#             when wapri = '09'
#             then 'production_operations_maintenance'
#             when wapri = '10'
#             then 'professional_services'
#             when wapri = '11'
#             then 'sales_purchasing_marketing'
#             when wapri = '12'
#             then 'quality_productivity_management'
```

```
#             when wapri = '13'
#             then 'teaching'
#             when wapri = '14'
#             then 'other_work_activity'
#             else null
#             end as wapri,  --Work activity spent most hours on in principal job - like dept
#         wasec,  --Work activity spent second most hours on in principal job
#         newbus, --within 5 years founded
#         case
#             when cast(nedtp as int) = 1 then 'self_employed_not_incorp'
#             when cast(nedtp as int) = 2 then 'self_employed_incorp'
#             when cast(nedtp as int) = 3 then 'private_for_profit_org'
#             when cast(nedtp as int) = 4 then 'private_not_for_profit_org'
#             when cast(nedtp as int) = 5 then 'local_government'
#             when cast(nedtp as int) = 6 then 'state_government'
#             when cast(nedtp as int) = 7 then 'military'
#             when cast(nedtp as int) = 8 then 'federal_government'
#             when cast(nedtp as int) = 9 then 'other_employer_type'
#         else null
#         end as employer_type,
#         case
#             when cast(emsecsm as int) = 1 then 'educational_institutution'
#             when cast(emsecsm as int) = 2 then 'government'
#             when cast(emsecsm as int) = 3 then 'business'
#         else null
#         end as sector, -- Employer sector --1: Educational Institution 2: Government 3: Business/Industry
#         case when
#             emsize = 1
#             then '10_or_fewer_employees'
#         when emsize = 2
#             then '11_to_24_employees'
#         when emsize = 3
#             then '25_to_99_employees'
#         when emsize = 4
#             then '100_to_499_employees'
#         when emsize = 5
#             then '500_to_999_employees'
#         when emsize = 6
#             then '1000_to_4999_employees'
#         when emsize = 7
#             then '5000_to_24999_employees'
#         when emsize = 8
#             then '25000+_employees'
#         else null
#         end as emsize,
```

```
#         case
#         when emsize = 1
#             then 10
#         when emsize = 2
#             then round(11 + (24-11)/2)
#         when emsize = 3
#             then round(25 + (99-25)/2)
#         when emsize = 4
#             then round(100 + (499-100)/2)
#         when emsize = 5
#             then round(500 + (999-500)/2)
#         when emsize = 6
#             then round(1000 + (4999-1000)/2)
#         when emsize = 7
#             then round(5000 + (24999-5000)/2)
#         when emsize = 8
#             then 25000
#         else null
#         end as emsize_midpoint,
#         case
#             when emrg = '01'
#                 then 'new_england'
#             when emrg = '02'
#                 then 'middle_atlantic'
#             when emrg = '03'
#                 then 'east_north_central'
#             when emrg = '04'
#                 then 'west_north_central'
#             when emrg = '05'
#                 then 'south_atlantic'
#             when emrg = '06'
#                 then 'east_south_central'
#             when emrg = '07'
#                 then 'west_south_central'
#             when emrg = '08'
#                 then 'mountain'
#             when emrg = '09'
#                 then 'pacific_us_territories'
#             when emrg = '10'
#                 then 'europe'
#             when emrg = '20'
#                 then 'asia'
#             when emrg = '33'
#                 then 'caribbean'
#             when emrg = '37'
#                 then 'south_america'
#             when emrg = '40'
```

```
#                 then 'africa'
#             else null
#             end as emrg,
#         case when
#             emst = '085' then 'new_england'
#             when emst = '086' then 'mid_atlantic'
#             when emst = '087' then 'east_north_central'
#             when emst = '088' then 'west_north_central'
#             when emst = '089' then 'south_atlantic'
#             when emst = '090' then 'east_south_central'
#             when emst = '091' then 'west_south_central'
#             when emst = '092' then 'mountain'
#             when emst = '093' then 'pacific_region'
#             when emst = '096' then 'not_specified'
#             else null
#         end as emst,
#         case
#             when cast(jobsatis as int) = 1
#                 then 'very_satisfied'
#             when cast(jobsatis as int) = 2
#                 then 'somewhat_satisfied'
#             when cast(jobsatis as int) = 3
#                 then 'somewhat_dissatisfied'
#             when cast(jobsatis as int) = 4
#                 then 'very_dissatisfied'
#             end as jobsatis,
#         case
#             when
#                 cast(wkslyr as int) = 98 --logical skip
#                 then 52
#             else cast(wkslyr as int)
#         end as wkslyr, --Number of weeks worked per year if less than 52 weeks
#         cast(refyr as int) - cast(strtyr as int) as years_at_job,
#         case
#             when
#                 wksyr = 1 --Was this salary based on a 52-week year, or less␣
 ↪than that? 1 = worked 52 weeks per year
#                 then 1
#             else 0
#         end as fiftytwo_wksyr,
#         cast(refyr as int) - cast(dgryr as int) as years_since_highest_degree,
#         cast(refyr as int) - cast(mryr as int) as␣
 ↪years_since_most_recent_degree,
#         cast(refyr as int) - cast(bayr as int) as years_since_first_ba,
#         cast(bsdgn as int) as bsdgn, --Number of bachelor's or higher degrees
#         cast(bayr as int) as bayr,
#         baind, --Indicator of whether respondent received a BA-level degree
```

```
#         hdcarn,
#         n2dgrmed, -- Field of study for highest degree - best code
#         ndgmebg, --Field of study for highest degree broad group - SE based
#         case
#         when n2dgrmed like '116%'
#             then 'computer_and_info_sciences'
#         when n2dgrmed like '128%'
#             then 'math'
#         when n2dgrmed like '216%'
#             then 'agricultural_sciences'
#         when n2dgrmed like '226%'
#             then 'biological_sciences'
#         when n2dgrmed like '236%'
#             then 'environmental_sciences'
#         when n2dgrmed like '318%'
#             then 'chemistry'
#         when n2dgrmed like '328%'
#             then 'geological_sciences'
#         when n2dgrmed like '338%'
#             then 'physics'
#         when n2dgrmed like '348%'
#             then 'other_physical_sciences'
#         when n2dgrmed like '416%'
#             then 'ag_economics'
#         when n2dgrmed like '419%'
#             then 'economics'
#         when n2dgrmed like '429%'
#             then 'political_sciences'
#         when n2dgrmed like '437%'
#             then 'ed_psychollology'
#         when n2dgrmed like '438%'
#             then 'psychollology'
#         when n2dgrmed like '449%'
#             then 'sociology_anthropology'
#         when n2dgrmed like '456%'
#             then 'ethnic_studies'
#         when n2dgrmed like '457%'
#             then 'linguistics'
#         when n2dgrmed like '458%'
#             then 'philosophy'
#         when n2dgrmed like '459%'
#             then 'geography'
#         when n2dgrmed like '517%'
#             then 'aerospace_eng'
#         when n2dgrmed like '527%'
#             then 'chem_eng'
#         when n2dgrmed like '537%'
```

```
#             then 'civil_arch_eng'
#         when n2dgrmed like '547%'
#             then 'computer_eng'
#         when n2dgrmed like '557%'
#             then 'industrial_eng'
#         when n2dgrmed like '567%'
#             then 'mechanical_eng'
#         when n2dgrmed like '577%'
#             then 'bio_eng'
#         when n2dgrmed like '617%'
#             then 'health_medical_sciences'
#         when n2dgrmed like '627%'
#             then 'specialized_teacher_education'
#         when n2dgrmed like '636%'
#             then 'computer_programming'
#         when n2dgrmed like '637%'
#             then 'other_eng'
#         when n2dgrmed like '6461%'
#             then 'architects'
#         when n2dgrmed like '6465%'
#             then 'acutuaries'
#         when n2dgrmed like '716%'
#             then 'business_management'
#         when n2dgrmed like '727%'
#             then 'education'
#         when n2dgrmed like '738%'
#             then 'theology'
#         when n2dgrmed like '739%'
#             then 'social_work'
#         when n2dgrmed like '746%'
#             then 'marketing'
#         when n2dgrmed like '757%'
#             then 'literature_language'
#         when n2dgrmed like '758%'
#             then 'liberal_arts'
#         when n2dgrmed like '7592%'
#             then 'history'
#         when n2dgrmed like '7594%'
#             then 'arts'
#         when n2dgrmed like '7666%'
#             then 'journalism_communications'
#         when n2dgrmed like '7666%'
#             then 'journalism_communications'
#         when n2dgrmed like '7668%'
#             then 'conservation'
#         when n2dgrmed like '7669%'
#             then 'criminal_justice'
```

```
#         when n2dgrmed like '7680%'
#             then 'home_econ'
#         when n2dgrmed like '7681%'
#             then 'law'
#         when n2dgrmed like '7683%'
#             then 'library_science'
#         when n2dgrmed like '7685%'
#             then 'parks_and_rec'
#         when n2dgrmed like '7690%'
#             then 'public_admin'
#         when n2dgrmed = '769950'
#             then 'other'
#         else null
#         end as highest_degree_field,
#         case
#             when cast(ndgmemg as int) = 1
#                 then 'computer_math_sciences'
#             when cast(ndgmemg as int) = 2
#                 then 'life_sciences'
#             when cast(ndgmemg as int) = 3
#                 then 'physical_related_sciences'
#             when cast(ndgmemg as int) = 4
#                 then 'social_related_sciences'
#             when cast(ndgmemg as int) = 5
#                 then 'engineering'
#             when cast(ndgmemg as int) = 6
#                 then 'S&E_related_fields'
#             when cast(ndgmemg as int) = 7
#                 then 'non_S&E_fields'
#             end as highest_general_degree_field,
#         ndgmeng, -- Field of study for highest degree minor group
#         case
#             when
#                 dgrdg = 1
#                 then 'Bachelors'
#             when
#                 dgrdg = 2
#                 then 'Masters'
#             when
#                 dgrdg = 3
#                 then 'Doctorate'
#             when
#                 dgrdg = 4
#                 then 'Professional'
#         end as highest_degree_type,
#         spnat, --Technical expertise required by spouse/partner job:␣
 ↪engineering, computer, math, natural sciences
```

```
#         spot, --Technical expertise required by spouse/partner job: other
#         spsoc, --Technical expertise required by spouse/partner Job: social␣
 ↪sciences
#         case
#             when spowk = 1
#                 then 'full_time'
#             when spowk = 2
#                 then 'part_time'
#             when spowk = 3
#                 then 'non_working'
#             else 'not_partnered'
#         end as partner_work,
#         marind,
#         case
#             when cast(marsta as int) = 1
#                 then 'married'
#             when cast(marsta as int) = 2
#                 then 'living_with_partner'
#             when cast(marsta as int) = 3
#                 then 'widowed'
#             when cast(marsta as int) = 4
#                 then 'separated'
#             when cast(marsta as int) = 5
#                 then 'divorced'
#             when cast(marsta as int) = 6
#                 then 'never_married'
#         end as marsta,
#         chlvin, --Children living in household indicator (all ages)
#         CASE
#             WHEN (CAST(marsta AS int) = 1 OR CAST(marsta AS int) = 2) AND␣
 ↪chlvin = 1 THEN 1
#             ELSE 0
#         END AS partnered_with_children,
#         ch6in,
#         ch1218in,
#         ch19in,
#         ch25in,
#         ch611in,
#         chun12,
#         case
#             when
#                 cast(chu2 as int) = 98
#                 then null
#             else cast(chu2 as int)
#         end as chu2,
#         case
#             when
```

```
#                 cast(ch25 as int) = 98
#                 then null
#             else cast(ch25 as int)
#         end as ch25,
#         case
#             when
#                 cast(ch6 as int) = 98
#                 then null
#             else cast(ch6 as int)
#         end as ch6,
#         case
#             when
#                 cast(ch611 as int) = 98
#                 then null
#             else cast(ch611 as int)
#         end as ch611,
#         case
#             when
#                 cast(ch1218 as int) = 98
#                 then null
#             else cast(ch1218 as int)
#         end as ch1218,
#         case
#             when
#                 cast(ch19 as int) = 98
#                 then null
#             else cast(ch19 as int)
#         end as ch19,
#         lfstat,
#         resplcus
#     from df
#     where df.lfstat = 1 --employed workforce
#     and df.resplcus = 1 --respondents in US
#     and baind = 1 --Indicator of whether respondent received a BA-level degree
#     and agegr < 67 -- retirement age
#     and salary > 15080 --minimum wage in dollars per year for full-time␣
 ↪workers
#     and earn > 15080
#     --and employer_type = 'private_for_profit_org'
#     and hrswk >=30 -- full time workers
#     and fiftytwo_wksyr = 1 --salary based on 52 weeks worked last year
#     and employer_type not in
#         (
#         'military',
#         'other_employer_type',
#         'self_employed_not_incorp',
#         'self_employed_incorp'
```

```
#         )
# """
# sql_query = jinja2.Template(raw_query).render(vars())
```

Descriptive Statistics

[ ]:
```
data.describe().T
```

1.2 Who is included in this sample?

- lfstat = 1 –employed workforce
- and resplcus = 1 –respondents in US
- and baind = 1 –Indicator of whether respondent received a BA-level degree
- and agegr < 67 – retirement age
- and salary > 15080 –minimum wage in dollars per year for full-time workers–
- and hrswk >=30 – full time workersand fiftytwo_wksyr = 1 –salary based on 52 weeks worked last year
- and employer_type not in ('military', 'other_employer_type', 'self_employed_not_incorp', 'self_employed_incorp').

Using the same filtering mechanisms from class, we are looking at U.S. based respondents who are under the retirement age, make over minimum wage, and work full-time. We are not looking at people who work in the military, other employer types, or self-employed individuals.

[ ]:
```
# import jinja2
# raw_query = """
#     select
#     count(*),
#     employer_type
#     from data
#     group by employer_type
# """
# sql_query = jinja2.Template(raw_query).render(vars())
```

[ ]:
```
data.loc[(data['employer_type'] == 'private_for_profit_org'), 'private_corp'] =␣
↪1
data.loc[data['private_corp'].isnull(), 'private_corp'] = 0
data['private_corp'].describe()
```

[ ]:
```
count    58207.000000
mean         0.611078
std          0.487510
min          0.000000
25%          0.000000
50%          1.000000
75%          1.000000
max          1.000000
```

```
Name: private_corp, dtype: float64
```

1.3 Exploratory Data Analysis

1.3.1 Hypothesis 1- Null: Traditional family structures (partnered with children and married) do not significantly impact the gender wage gap or gender representation in management careers

****

_Hypothesis 1:___ (Multiple Regression and Chi-Squared Test) _

_Null:___ Traditional family structures (partnered with children and married) do not significantly impact the gender wage gap and gender representation in management careers. _

_Alternative: ___**Traditional family structures (partnered with children and married) have a significant impact on the gender wage gap or gender representation in management careers._

1.1 Average Salaries by Family Structure for Management Roles This section calculates the **average salaries** for individuals in **management roles** across **sectors** (business, educational institutions, government), grouped by **family structure** ('marind' = married, 'chlvin' = has children). So we visualize whether **traditional family structures (married with children)** influence salaries across different sectors, addressing Hypothesis 1 regarding the gender wage gap.

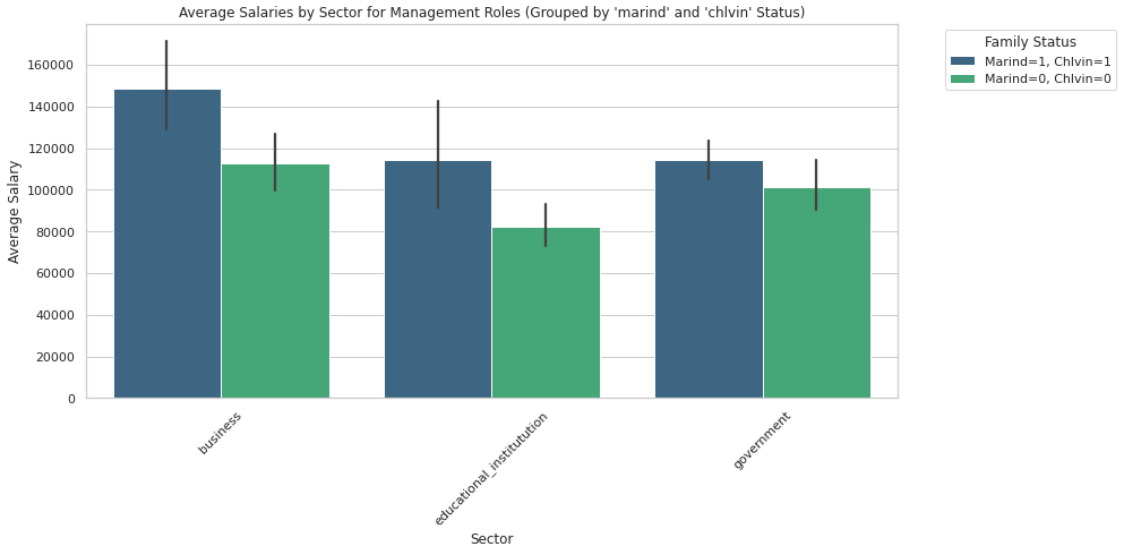
[ ]:
```
#create a dataframe of average salaries (mina)
average_salaries = (
    data.groupby(['principal_job', 'leadership_job_titles', 'sector', 'marind',␣
 ↪'chlvin'])
        .agg(
            number_of_respondents=('salary', 'size'),
            average_salary=('salary', 'mean'),
            average_earnings=('earn', 'mean')
        )
        .reset_index()
)

#get the percent female
percent_female = (
    data.groupby(['principal_job', 'leadership_job_titles', 'sector', 'marind',␣
 ↪'chlvin'])
        .agg(percent_female=('gender', 'mean'))
        .reset_index()
)

#merge the average salary and percent female tables
result = (
    average_salaries.merge(percent_female, on=['principal_job',␣
 ↪'leadership_job_titles', 'sector', 'marind', 'chlvin'])
        .assign(
```

```
            average_salary=lambda df: df['average_salary'].round(2),
            average_earnings=lambda df: df['average_earnings'].round(2),
            percent_female=lambda df: df['percent_female'].round(2)
        )
        .sort_values(['principal_job', 'sector', 'marind', 'chlvin'])
)

#show result
result[[
    'number_of_respondents',
    'principal_job',
    'sector',
    'leadership_job_titles',
    'marind',
    'chlvin',
    'average_salary',
    'average_earnings',
    'percent_female',
]]
```

[ ]:
```
#combine the subsets into one df with a label for family status (mina)
subset_marind_chlvin_1 = result[(result['marind'] == 1) & (result['chlvin'] ==␣
 ↪1)].copy()
subset_marind_chlvin_1['status'] = 'Marind=1, Chlvin=1'

subset_marind_chlvin_0 = result[(result['marind'] == 0) & (result['chlvin'] ==␣
 ↪0)].copy()
subset_marind_chlvin_0['status'] = 'Marind=0, Chlvin=0'

combined_data = pd.concat([subset_marind_chlvin_1, subset_marind_chlvin_0])

#sns theme
sns.set_theme(style="whitegrid")

#bar chart for average salaries
plt.figure(figsize=(14, 7))
sns.barplot(
    data=combined_data,
    x="sector",
    y="average_salary",
    hue="status",
    palette="viridis"
)
plt.title("Average Salaries by Sector for Management Roles (Grouped by 'marind'␣
 ↪and 'chlvin' Status)")
plt.ylabel("Average Salary")
plt.xlabel("Sector")
```

```
plt.legend(title="Family Status", bbox_to_anchor=(1.05, 1), loc='upper left')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

[ ]:



- In the ,**business**, sector, individuals from traditional family structures (Marind=1, Chlvin=1) have significantly higher average salaries compared to those who are unmarried and without children (Marind=0, Chlvin=0).
- In the ,**education**, and ,**government**, sectors, the gap is smaller but still evident, suggesting that traditional family structures may confer advantages.

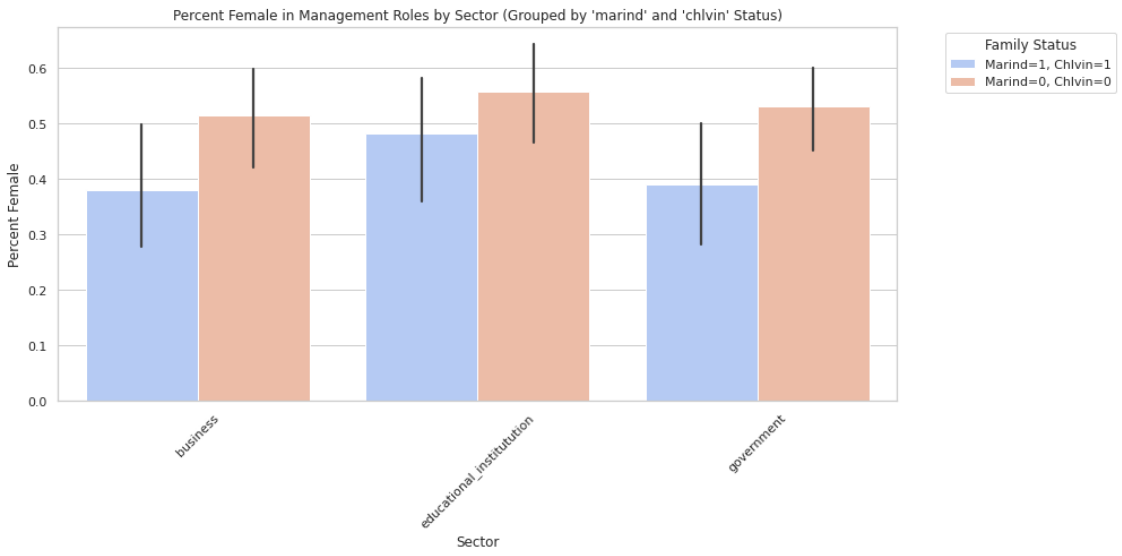
This figure supports the **alternative hypothesis** for Hypothesis 1, indicating that family structure significantly impacts salaries in management careers.

1.2 Percent Female in Management Roles by Family Structure The following part analyzes the **proportion of females** in **management roles** across **sectors and family structures**. By comparing these proportions, it helps assess whether female representation might correlate with the gender wage gap and whether traditional family structures influence this representation.

[ ]:
```
#Bar Chart: Percent Female (zhilie)
plt.figure(figsize=(14, 7))
sns.barplot(
    data=combined_data,
    x="sector",
    y="percent_female",
    hue="status",
    palette="coolwarm"
)
```

```
plt.title("Percent Female in Management Roles by Sector (Grouped by 'marind'␣
 ↪and 'chlvin' Status)")
plt.ylabel("Percent Female")
plt.xlabel("Sector")
plt.legend(title="Family Status", bbox_to_anchor=(1.05, 1), loc='upper left')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

[ ]:



- The ,**education**, sector has the highest proportion of females in management roles, particularly among Unmarried and childless individuals (Marind=0, Chlvin=0). They have slightly higher female representation(55%) compared to their married counterparts with children (Marind=1, Chlvin=1) (48%).
- The ,**business**, and ,**government**, sectors have lower female representation, with married individuals with children (Marind=1, Chlvin=1) having even less representation.

Based on the visualization here, we may want to **support the alternative hypothesis** that the traditional family structures limit women's access and representation to management roles.

1.3 Distribution of Salary by Sector, Gender, and Family Status In this section, we introduce a box plot to visualize the distribution of **salaries** across **sectors**, grouped by **gender** and **family structure**. It highlights salary differences at different levels, such as the median and outliers, to assess the intersectionality of gender and family structure.

[ ]:
```
average_salaries = (
    data.groupby(['principal_job', 'leadership_job_titles', 'sector', 'marind',␣
 ↪'chlvin', 'gender'])
        .agg(
            number_of_respondents=('salary', 'size'),
```

```
            average_salary=('salary', 'mean'),
            average_earnings=('earn', 'mean')
        )
        .reset_index()
)
```

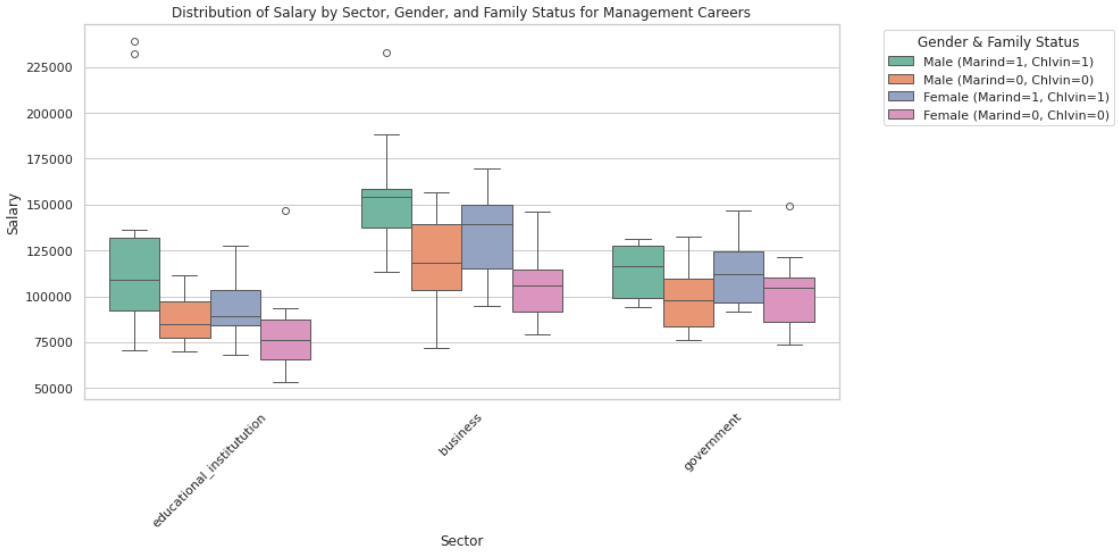
[ ]:
```
average_salaries.head(5)
```

[ ]:
```
#pivot the data to have male and female salaries in separate columns (mina)
gender_pivot = average_salaries.pivot_table(
    index=['principal_job', 'leadership_job_titles', 'sector', 'marind',␣
 ↪'chlvin'],
    columns='gender',
    values='average_salary'
).reset_index()

#rename columns
gender_pivot.columns.name = None  # Remove hierarchical column names
gender_pivot = gender_pivot.rename(columns={0: 'male_salary', 1:␣
 ↪'female_salary'})  # Assuming 0 = Male, 1 = Female

#calculate gender wage gap
gender_pivot['gender_wage_gap'] = (
    (gender_pivot['male_salary'] - gender_pivot['female_salary']) /␣
 ↪gender_pivot['male_salary']
) * 100

#sort by gender wage gap
gender_pivot = gender_pivot.sort_values(by='gender_wage_gap', ascending=False)
```

[ ]:
```
#filter for family status (mina)
filtered_gender_pivot = gender_pivot[
    ((gender_pivot['marind'] == 0) & (gender_pivot['chlvin'] == 0)) |
    ((gender_pivot['marind'] == 1) & (gender_pivot['chlvin'] == 1))
].copy()

#combine for family status
filtered_gender_pivot['family_status'] = filtered_gender_pivot.apply(
    lambda row: 'Marind=1, Chlvin=1' if (row['marind'] == 1 and row['chlvin']␣
 ↪== 1) else 'Marind=0, Chlvin=0',
    axis=1
)

#melt to get male and female salaries for box plot
salary_distribution = filtered_gender_pivot.melt(
    id_vars=['sector', 'family_status'],
    value_vars=['male_salary', 'female_salary'],
```

```
    var_name='gender',
    value_name='salary'
)

#re-label
salary_distribution['gender'] = salary_distribution['gender'].replace({
    'male_salary': 'Male',
    'female_salary': 'Female'
})

#gender_family_status label for use in box plot
salary_distribution['gender_family_status'] = salary_distribution['gender'] + ' ␣
 ↪(' + salary_distribution['family_status'] + ')'

#create a box plot
plt.figure(figsize=(14, 7))
sns.boxplot(
    data=salary_distribution,
    x='sector',
    y='salary',
    hue='gender_family_status',
    palette='Set2'
)
plt.title("Distribution of Salary by Sector, Gender, and Family Status for␣
 ↪Management Careers")
plt.ylabel("Salary")
plt.xlabel("Sector")
plt.legend(title="Gender & Family Status", bbox_to_anchor=(1.05, 1), loc='upper␣
 ↪left')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

[ ]:



Educational Institutions:

- **Men in traditional family structures (Male, Marind=1, Chlvin=1)**, have the highest median salary in this sector, with relatively narrow salary variability.in t
- **Women in traditional family structure (Female, Marind=1, Chlvin=1):**, Although their median salary is higher than unmarried women, it is still significantly lower than that of their male counterparts.
- **Unmarried Individuals (Marind=0, Chlvin=0):**, Both men and women in this group earn significantly lower salaries, with unmarried women earning the least.

Business Sector:

- **Men in traditional family structures (Male, Marind=1, Chlvin=1)** ,still have the highest median salary overall, showcasing the largest gap relative to women in the same family structure.
- **Women in traditional family structure(Female, Marind=1, Chlvin=1):**, Their salary distribution is wider, indicating greater variability, but they earn substantially less than men.
- **Unmarried Individuals (Marind=0, Chlvin=0):**, Salaries are lower for both genders, but women in this group experience the largest wage penalty.

Government Sector:

- **Individuals in Traditional family structure(Marind=1, Chlvin=1):**, Both genders have similar median salaries and variability, with nearly no wage gap.
- **Unmarried Individuals (Marind=0, Chlvin=0):**, Both genders face obviously lower median salaries compared to traditional family structure, though the gap between men and women in this category is less pronounced than in other sectors.

Except for the government sector, the pronounced gender wage gap exists across sectors, with males always earning more than females in both family structures. With this gap existing, unmarried Individuals (Marind=0, Chlvin=0)** **earn consistently lower salaries across all three sectors**

compared to individuals in traditional family structures (Marind=1, Chlvin=1). The results strongly support the alternative hypothesis**, as traditional family structures (married with children) significantly impact the gender wage gap.

1.4 Chi-Square Test

- between ,**family structure**, (marital and parental status: marind and chlvin) and ,**gender representation**, (female vs. non-female) across ,**sectors**, (business, educational institutions, government)

[ ]:
```
import numpy as np
import scipy.stats as stats
import pandas as pd


#create a contingency table for family structure and female representation
contingency_table = (
    data.groupby(["sector", "marind", "chlvin"])
    .agg(
        total_female=("gender", lambda x: (x == "female").sum()),
        total_non_female=("gender", lambda x: (x != "female").sum()),
    )
    .reset_index()
)

#combine married and child status into a single label
contingency_table["family_status"] = contingency_table.apply(
    lambda row: "Marind=1, Chlvin=1"
    if (row["marind"] == 1 and row["chlvin"] == 1)
    else "Marind=0, Chlvin=0",
    axis=1,
)

#pivot the table to create the contingency table for Chi-Square test
pivot_table = pd.pivot_table(
    contingency_table,
    values=["total_female", "total_non_female"],
    index="family_status",
    columns="sector",
    aggfunc="sum",
    fill_value=0  # Replace missing values with 0
)

#flatten the multi-index columns for better readability
pivot_table.columns = [f"{col[1]}_{col[0]}" for col in pivot_table.columns]

#convert pivot table to observed frequency matrix
observed = pivot_table.values
```

```
#check for zero rows/columns in observed data
row_sums = observed.sum(axis=1)
col_sums = observed.sum(axis=0)

#remove rows/columns with zero totals
observed = observed[row_sums > 0, :]
observed = observed[:, col_sums > 0]

#perform Chi-Squared test
chi2, p, dof, expected = stats.chi2_contingency(observed)

#display results
print("Chi-Squared Test Results")
print(f"Chi-Squared Value: {chi2:.2f}")
print(f"p-value: {p:.4f}")
print(f"Degrees of Freedom: {dof}")
print("\nExpected Frequencies:")
print(expected)
```

[ ]:
```
Chi-Squared Test Results
Chi-Squared Value: 6.98
p-value: 0.0306
Degrees of Freedom: 2

Expected Frequencies:
[[24789.91203807  5303.6621025   5268.42585943]
 [16015.08796193  3426.3378975   3403.57414057]]
```

- **Chi-Square Value**,: The Chi-Squared value (6.98) reflects the magnitude of the difference between observed and expected frequencies. Higher values indicate a greater deviation from the null hypothesis.
- **Degrees of Freedom**,: The degrees of freedom (2) correspond to a contingency table with 3 sectors (columns) and 2 family status categories (rows): df = (3-1) * (2-1) = 2
- **p-value**,: The p-value (0.0306) is less than the typical threshold of 0.05. This indicates a statistically significant relationship between ,**family status**, and ,**female representation**,.

Since the **p-value (0.0306)** is **less than 0.05**, we **reject the null hypothesis**. This means there is **evidence of a significant relationship** between family status and female representation in management roles across sectors.

1.5 Analysis for Control Variables: Family Structure and Gender Impact

[ ]:
```
import numpy as np
import scipy.stats as stats
import pandas as pd

#ensure relevant columns exist in the data
```

```
required_columns = ["sector", "marind", "chlvin", "gender", "ch6in", "ch611in",␣
 ↪"ch1218in", "ch19in", "emsize", "hrswk", "agegr", "newbus"]
missing_columns = [col for col in required_columns if col not in data.columns]
if missing_columns:
    raise ValueError(f"Missing columns in data: {missing_columns}")

#create hour bins for `hrswk` with 10-hour intervals
hour_bins = [0, 30, 40, 50, 60, 70, 80, 90, 100]  # Define bins: 0-30, 30-40, ..
 ↪., 90+
hour_labels = [
    "0-30 hours",
    "30-40 hours",
    "40-50 hours",
    "50-60 hours",
    "60-70 hours",
    "70-80 hours",
    "80-90 hours",
    "90+ hours",
]
data["hour_range"] = pd.cut(data["hrswk"], bins=hour_bins, labels=hour_labels,␣
 ↪right=False)

#function to perform Chi-Square test within a subgroup
def chi_square_test_within_group(subset_data, group_name):
    #create a contingency table
    contingency_table = (
        subset_data.groupby(["sector", "marind", "chlvin"])
        .agg(
            total_female=("gender", lambda x: (x == "female").sum()),
            total_non_female=("gender", lambda x: (x != "female").sum()),
        )
        .reset_index()
    )

    #combine married and child status into a single label
    contingency_table["family_status"] = contingency_table.apply(
        lambda row: "Marind=1, Chlvin=1"
        if (row["marind"] == 1 and row["chlvin"] == 1)
        else "Marind=0, Chlvin=0",
        axis=1,
    )

    #pivot table for Chi-Square test
    pivot_table = pd.pivot_table(
        contingency_table,
        values=["total_female", "total_non_female"],
        index="family_status",
```

```
        columns="sector",
        aggfunc="sum",
        fill_value=0
    )

    #flatten multi-index columns
    pivot_table.columns = [f"{col[1]}_{col[0]}" for col in pivot_table.columns]
    observed = pivot_table.values

    #check for zero rows/columns and remove them
    row_sums = observed.sum(axis=1)
    col_sums = observed.sum(axis=0)
    filtered_observed = observed[row_sums > 0, :]
    filtered_observed = filtered_observed[:, col_sums > 0]

    #perform Chi-Square test
    chi2, p, dof, expected = stats.chi2_contingency(filtered_observed)

    #append summarized results
    return {
        "group": group_name,
        "chi2": chi2,
        "p_value": p,
        "dof": dof,
        "conclusion": "Significant" if p < 0.05 else "Not Significant"
    }

#define children age groups as a control variable
children_age_vars = {
    "Under 6": "ch6in",
    "6-11": "ch611in",
    "12-18": "ch1218in",
    "19+": "ch19in"
}

#iterate through all control variables
control_vars = {
    "Children Age Group": children_age_vars,  # Nested control variable
    "Employer Size (Large vs. Small)": "emsize",
    "Hours Worked Per Week (Grouped)": "hour_range",  # Use 10-hour bins for␣
 ↪`hrswk`
    "Age Group": "agegr",
    "New Business": "newbus"
}

#collect and group results
grouped_results = {}
```

```
for control_label, control_var in control_vars.items():
    results = []
    if control_label == "Children Age Group":
        # Handle nested children age groups
        for age_label, age_var in control_var.items():
            for value in data[age_var].dropna().unique():
                subset_data = data[data[age_var] == value]
                result = chi_square_test_within_group(subset_data,␣
↪f"{age_label} = {value}")
                results.append(result)
        grouped_results[control_label] = pd.DataFrame(results).
↪sort_values("group")
    else:
        # Handle all other control variables
        for value in data[control_var].dropna().unique():
            subset_data = data[data[control_var] == value]
            result = chi_square_test_within_group(subset_data,␣
↪f"{control_label} = {value}")
            results.append(result)
        grouped_results[control_label] = pd.DataFrame(results).
↪sort_values("group")

#display results by each control variable
for control_label, results_df in grouped_results.items():
    print(f"\n=== Results for {control_label} ===\n")
    print(results_df)
```

[ ]:

```
=== Results for Children Age Group ===

           group       chi2       p_value  dof   conclusion
5    12-18 = 0.0  67.078613  2.716842e-15    2  Significant
4    12-18 = 1.0  27.157297  1.267266e-06    2  Significant
6      19+ = 0.0  83.185714  8.638768e-19    2  Significant
7      19+ = 1.0  14.491846  7.130757e-04    2  Significant
2     6-11 = 0.0  82.997026  9.493466e-19    2  Significant
3     6-11 = 1.0  21.982431  1.684906e-05    2  Significant
1  Under 6 = 0.0  54.177996  1.719482e-12    2  Significant
0  Under 6 = 1.0  15.581470  4.135489e-04    2  Significant

=== Results for Employer Size (Large vs. Small) ===

                                              group       chi2       p_value  \
5  Employer Size (Large vs. Small) = 1000_to_4999…   3.151216  2.068818e-01
4  Employer Size (Large vs. Small) = 100_to_499_e…  42.568481  5.706525e-10
7  Employer Size (Large vs. Small) = 10_or_fewer_…   5.246276  7.257475e-02
```

```
6  Employer Size (Large vs. Small) = 11_to_24_emp…   1.739223  4.191144e-01
1  Employer Size (Large vs. Small) = 25000+_emplo…  35.751671  1.724341e-08
2  Employer Size (Large vs. Small) = 25_to_99_emp…   1.672113  4.334162e-01
0  Employer Size (Large vs. Small) = 5000_to_2499…  31.037882  1.820579e-07
3  Employer Size (Large vs. Small) = 500_to_999_e…   9.436573  8.930466e-03

   dof       conclusion
5    2  Not Significant
4    2      Significant
7    2  Not Significant
6    2  Not Significant
1    2      Significant
2    2  Not Significant
0    2      Significant
3    2      Significant

=== Results for Hours Worked Per Week (Grouped) ===

                                          group       chi2   p_value  dof  \
2  Hours Worked Per Week (Grouped) = 30-40 hours   4.570628  0.101742    2
0  Hours Worked Per Week (Grouped) = 40-50 hours   5.692533  0.058061    2
3  Hours Worked Per Week (Grouped) = 50-60 hours   2.528112  0.282506    2
1  Hours Worked Per Week (Grouped) = 60-70 hours  24.207168  0.000006    2
5  Hours Worked Per Week (Grouped) = 70-80 hours   0.224289  0.893915    2
4  Hours Worked Per Week (Grouped) = 80-90 hours   3.119362  0.210203    2
6    Hours Worked Per Week (Grouped) = 90+ hours   1.041931  0.593947    2

        conclusion
2  Not Significant
0  Not Significant
3  Not Significant
1      Significant
5  Not Significant
4  Not Significant
6  Not Significant

=== Results for Age Group ===

            group       chi2       p_value  dof       conclusion
9  Age Group = 20   1.776180  4.114409e-01    2  Not Significant
8  Age Group = 25   3.041542  2.185433e-01    2  Not Significant
1  Age Group = 30   4.881815  8.708180e-02    2  Not Significant
3  Age Group = 35  20.192979  4.122402e-05    2      Significant
0  Age Group = 40  36.736804  1.053669e-08    2      Significant
6  Age Group = 45  30.299066  2.634155e-07    2      Significant
5  Age Group = 50  11.656830  2.942738e-03    2      Significant
2  Age Group = 55   3.265345  1.954067e-01    2  Not Significant
```

```
4  Age Group = 60   1.478268  4.775274e-01    2  Not Significant
7  Age Group = 65   1.541924  4.625680e-01    2  Not Significant

=== Results for New Business ===

                group      chi2   p_value  dof       conclusion
1  New Business = 0.0  8.127372  0.017186    2      Significant
0  New Business = 1.0  1.777399  0.411190    2  Not Significant
```

Children Age Group as a Control Variable:

- Significant results across all child age groups (Under 6, 6–11, 12–18, and 19+), both for the presence (=1) and absence (=0) of children.
- Chi-squared values are consistently high, indicating a strong association between family structure and gender representation across all child age groups.
- The presence of children in any age group significantly moderates the relationship between family structure and gender representation. This underscores the dual burden of professional and caregiving responsibilities often faced by women, further limiting their career growth and earning potential.
- Reject the null hypothesis

Employer Size (Large vs. Small) as a Control Variable:

- Significant results for large employers (e.g., 25000+ and 5000_to_24999 employees). Not significant for small employers (10_or_fewer and 11_to_24 employees).
- Larger employers show a stronger association between family structure and gender representation compared to smaller employers. Larger organizations may have more rigid structures, amplifying biases against women, particularly those with traditional family responsibilities. Conversely, smaller companies may offer more flexibility, reducing the observed impact of family structure on gender outcomes.
- Partially reject the null hypothesis [for large employers]

Hours Worked Per Week as a Control Variable:

- Significant results for individuals working 60–70 hours per week. Not significant for other hour ranges (30–60 hours or 90+ hours).
- The 60–70 hour range aligns with expectations for high-level management roles, which may disproportionately benefit men in traditional family structures. But in general, hours worked per week do not have a strong association between family structure and gender representation.
- Partially reject the null hypothesis [for individuals working 60–70 hours per week]

Age Group as a Control Variable:

- Significant results for mid-career age groups (e.g., 35–50 years).Not significant for younger (e.g., 20–30 years) or older (e.g., 60+ years) age groups.
- Family structure has the strongest impact during mid-career stages, a period when many individuals balance peak caregiving responsibilities with career advancement. Younger individuals may not yet face family-related challenges, while older individuals may have fewer caregiving responsibilities, reducing the observed impact of family structure on gender outcomes. This reinforces the notion that mid-career is a critical period for addressing gender

equity.

- Partially reject the null hypothesis [for mid-age groups]

New Business as a Control Variable:

- Significant results for established businesses (New Business = 0.0). Not significant for new businesses (New Business = 1.0).
- Established businesses likely reflect traditional workplace norms and biases, amplifying the impact of family structure on gender representation. New businesses may adopt more inclusive policies or lack entrenched norms, mitigating the impact of family structure. This suggests a potential pathway for reducing gender disparities through organizational culture change.
- Partially reject the null hypothesis [for established businesses]

Family structure consistently interacts with gender representation and outcomes across almost all control variables. The moderating effects are strongest in partnered with children of all ages, Mid-career stages (Age Group = 35–50 years); High-hour workloads (60–70 hours); and Large, established organizations. In most scenarios focusing on the different control variables, **the null hypothesis is rejected**, indicating that traditional family structures (married with children) significantly impact the gender representation gap in management careers.

1.6 Multiple Regression

[ ]:
```
import pandas as pd
import numpy as np
import statsmodels.api as sm
import statsmodels.formula.api as smf

#prepare the dataset (mina)
regression_data = data[[
    'salary', 'gender', 'marind', 'chlvin', 'leadership_job_titles', 'sector'
]].dropna()

#create interaction terms
regression_data['gender_marind_interaction'] = regression_data['gender'] *␣
 ↪regression_data['marind']
regression_data['gender_chlvin_interaction'] = regression_data['gender'] *␣
 ↪regression_data['chlvin']

#encode categorical variables
regression_data = pd.get_dummies(
    regression_data,
    columns=['sector', 'leadership_job_titles'],
    drop_first=True
)

#prepare predictors for the formula
predictors = regression_data.columns.difference(['salary'])
formula = "salary ~ gender + marind + chlvin + gender_marind_interaction +␣
 ↪gender_chlvin_interaction + " + " + ".join(predictors)
```

```
#fit the regression model
model = smf.ols(formula=formula, data=regression_data).fit()

#display the summary of the regression results
print(model.summary())
```

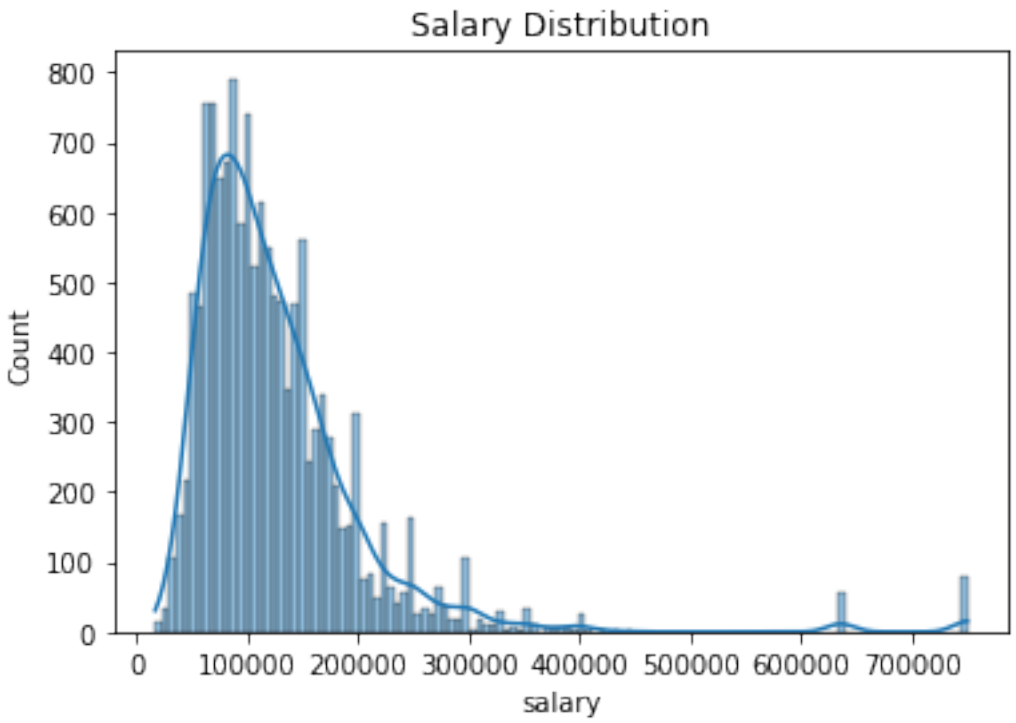
[ ]:
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 salary   R-squared:                       0.181
Model:                            OLS   Adj. R-squared:                  0.180
Method:                 Least Squares   F-statistic:                     189.4
Date:                Tue, 10 Dec 2024   Prob (F-statistic):               0.00
Time:                        11:07:36   Log-Likelihood:            -1.7419e+05
No. Observations:               13762   AIC:                         3.484e+05
Df Residuals:                   13745   BIC:                         3.485e+05
Df Model:                          16
Covariance Type:            nonrobust
==============================================================================
====================================
                                                    coef    std err
t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
------------------------------------
Intercept                                        1.737e+05   2449.556
70.897      0.000    1.69e+05    1.78e+05
gender                                          -8465.6634   2396.027
-3.533      0.000   -1.32e+04   -3769.123
marind                                           2.699e+04   2281.116
11.830      0.000    2.25e+04    3.15e+04
chlvin                                           1.051e+04   1968.244
5.342       0.000    6656.537    1.44e+04
gender_marind_interaction                       -1.218e+04   3148.149
-3.868      0.000   -1.83e+04   -6004.893
gender_chlvin_interaction                       -4723.6320   2906.033
-1.625      0.104   -1.04e+04     972.590
leadership_job_titles_computer_it_managers      -3.104e+04   3975.114
-7.808      0.000   -3.88e+04   -2.32e+04
leadership_job_titles_ed_administrators         -5.094e+04   5448.156
-9.351      0.000   -6.16e+04   -4.03e+04
leadership_job_titles_eng_managers              -5.036e+04   2833.280
-17.774      0.000   -5.59e+04   -4.48e+04
leadership_job_titles_financial_mgmt            -7.565e+04   2449.197
-30.888      0.000   -8.05e+04   -7.09e+04
leadership_job_titles_medical_health_manager    -5.491e+04   4208.226
-13.049      0.000   -6.32e+04   -4.67e+04
leadership_job_titles_natural_science_manager    -5.483e+04   4777.165
```

```
-11.478      0.000   -6.42e+04   -4.55e+04
leadership_job_titles_other_management_occupations -7.346e+04   2110.946
-34.798      0.000   -7.76e+04   -6.93e+04
leadership_job_titles_other_mid_level_manager      -5.653e+04   2486.870
-22.732      0.000   -6.14e+04   -5.17e+04
leadership_job_titles_personnel_mgmt                -8.89e+04   3250.619
-27.349      0.000   -9.53e+04   -8.25e+04
sector_educational_institututution                 -2.871e+04   2795.539
-10.270      0.000   -3.42e+04   -2.32e+04
sector_government                                  -2.146e+04   1823.065
-11.769      0.000    -2.5e+04   -1.79e+04
==============================================================================
Omnibus:                    11197.367   Durbin-Watson:                   1.867
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           357331.490
Skew:                           3.743   Prob(JB):                         0.00
Kurtosis:                      26.814   Cond. No.                         15.9
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

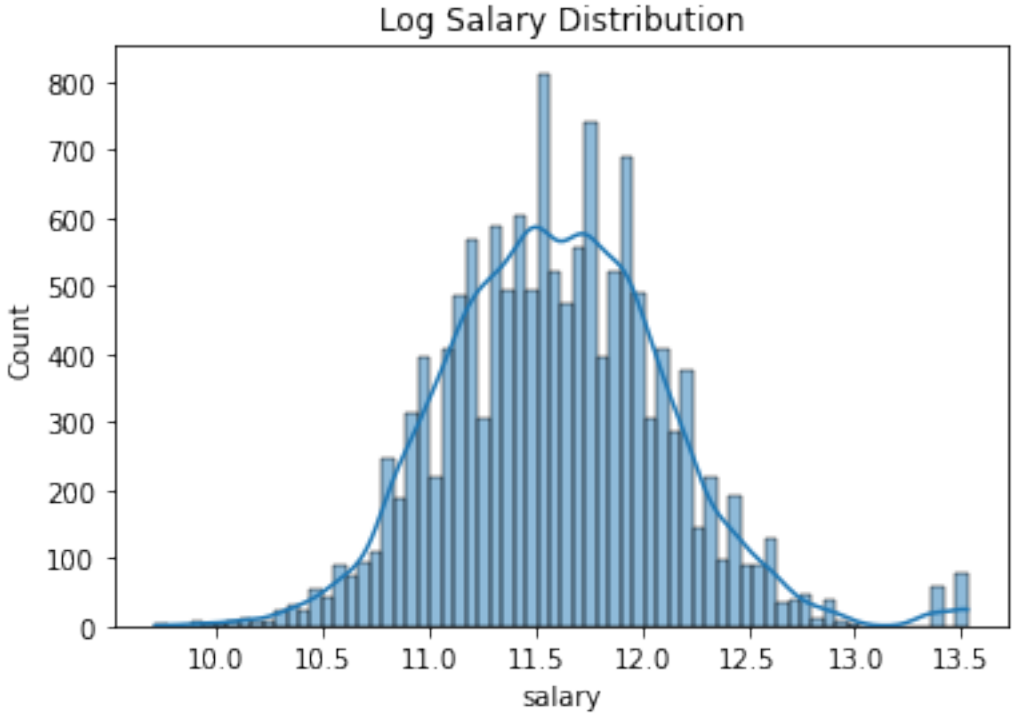
Before we do analysis, we check whether the distribution of salary data is skewed: If Skewness $>$ 1, it indicates severe right skewness, in which case **logarithmic transformation** of the data **is needed** to improve the performance of the regression model.

[ ]:
```python
import matplotlib.pyplot as plt
import seaborn as sns

sns.histplot(regression_data['salary'], kde=True)
plt.title('Salary Distribution')
plt.show()

sns.histplot(np.log(regression_data['salary']), kde=True)
plt.title('Log Salary Distribution')
plt.show()

print("Salary skewness:", regression_data['salary'].skew())
print("Log Salary skewness:", np.log(regression_data['salary']).skew())
```

[ ]:


Salary Distribution
Count
salary
0
100
200
300
400
500
600
700
800
0
100000
200000
300000
400000
500000
600000
700000

[ ]:


Log Salary Distribution
Count
salary
0
100
200
300
400
500
600
700
800
10.0
10.5
11.0
11.5
12.0
12.5
13.0
13.5

[ ]:
```
Salary skewness: 3.68561499050421
Log Salary skewness: 0.34898072637792854
```

From the graph we can interpret that, the skewed distribution may cause the model residuals to be inconsistent with the normal distribution hypothesis, thus affecting the accuracy of the p-value. High skewness amplifies the influence of outliers and makes the model results unstable. So below we decide to analyze the **'log salary' instead of 'salary'** to ensure the performance of model**.**

[ ]:
```
import pandas as pd
import numpy as np
import statsmodels.api as sm
import statsmodels.formula.api as smf

# Prepare the dataset (zhilie)
regression_data = data[[
    'salary', 'gender', 'marind', 'chlvin', 'leadership_job_titles', 'sector'
]].dropna()

# Ensure salary is positive before applying log transformation
regression_data = regression_data[regression_data['salary'] > 0]

# Create log-transformed salary
regression_data['log_salary'] = np.log(regression_data['salary'])

# Create interaction terms
regression_data['gender_marind_interaction'] = regression_data['gender'] *␣
 ↪regression_data['marind']
regression_data['gender_chlvin_interaction'] = regression_data['gender'] *␣
 ↪regression_data['chlvin']

# Encode categorical variables
regression_data = pd.get_dummies(
    regression_data,
    columns=['sector', 'leadership_job_titles'],
    drop_first=True
)

# Prepare predictors for the formula
predictors = regression_data.columns.difference(['salary', 'log_salary'])
formula = "log_salary ~ gender + marind + chlvin + gender_marind_interaction +␣
 ↪gender_chlvin_interaction + " + " + ".join(predictors)

# Fit the regression model
model = smf.ols(formula=formula, data=regression_data).fit()

# Display the summary of the regression results
```

```
print(model.summary())
```

[ ]:
```
                            OLS Regression Results
==============================================================================
Dep. Variable:             log_salary   R-squared:                       0.238
Model:                            OLS   Adj. R-squared:                  0.237
Method:                 Least Squares   F-statistic:                     268.0
Date:                Tue, 10 Dec 2024   Prob (F-statistic):               0.00
Time:                        11:09:59   Log-Likelihood:                -8621.7
No. Observations:               13762   AIC:                         1.728e+04
Df Residuals:                   13745   BIC:                         1.741e+04
Df Model:                          16
Covariance Type:            nonrobust
==============================================================================
                                                    coef    std err
t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
------------------------------------
Intercept                                        11.8405      0.015
811.323      0.000      11.812      11.869
gender                                           -0.0832      0.014
-5.831      0.000      -0.111      -0.055
marind                                            0.2126      0.014
15.645      0.000       0.186       0.239
chlvin                                            0.0723      0.012
6.166      0.000       0.049       0.095
gender_marind_interaction                        -0.0714      0.019
-3.809      0.000      -0.108      -0.035
gender_chlvin_interaction                        -0.0154      0.017
-0.889      0.374      -0.049       0.019
leadership_job_titles_computer_it_managers       -0.0883      0.024
-3.728      0.000      -0.135      -0.042
leadership_job_titles_ed_administrators          -0.1860      0.032
-5.729      0.000      -0.250      -0.122
leadership_job_titles_eng_managers               -0.2019      0.017
-11.958      0.000      -0.235      -0.169
leadership_job_titles_financial_mgmt             -0.4736      0.015
-32.459      0.000      -0.502      -0.445
leadership_job_titles_medical_health_manager     -0.2919      0.025
-11.642      0.000      -0.341      -0.243
leadership_job_titles_natural_science_manager    -0.2679      0.028
-9.414      0.000      -0.324      -0.212
leadership_job_titles_other_management_occupations  -0.4229      0.013
-33.624      0.000      -0.448      -0.398
leadership_job_titles_other_mid_level_manager    -0.2912      0.015
-19.653      0.000      -0.320      -0.262
```

```
leadership_job_titles_personnel_mgmt                    -0.6016      0.019
-31.066      0.000      -0.640      -0.564
sector_educational_institutution                        -0.2690      0.017
-16.149      0.000      -0.302      -0.236
sector_government                                       -0.1154      0.011
-10.623      0.000      -0.137      -0.094
==============================================================================
Omnibus:                      499.986   Durbin-Watson:                   1.808
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              935.921
Skew:                           0.284   Prob(JB):                    5.85e-204
Kurtosis:                       4.144   Cond. No.                         15.9
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

Gender Wage Gap [**gender**]: Women earn 8.32% less than men on average (significant).

Marital Status [**marind**]: Married individuals earn 21.26% more than unmarried individuals (significant).

[**gender_marind_interaction**]: Being married further reduces women's salary advantage by 7.14% compared to men (significant). Marriage significantly amplifies the gender wage gap, as its positive effect on salary benefits men more than women.

Children [chlvin]: Having children increases salary by 7.23% (significant).

[**gender_chlvin_interaction**]: The impact of having children on the gender wage gap is not significant, the p-value of 0.374>0.05, suggests that having children does not significantly exacerbate the gender wage gap.

Leadership Roles: Women are likely underrepresented in higher-paying leadership roles due to structural barriers: For example, roles such as **computer/IT managers** and **education administrators** show significant negative salary coefficients, highlighting lower pay within these leadership categories. Women in management roles within lower-paying **sectors (e.g., education)** face compounded disadvantages, earning even less compared to men.

Sectors: Lower-paying sectors like **education** disproportionately affect women, given their higher representation in such fields. This reinforces systemic issues where women are concentrated in industries and roles that pay less overall.

Marriage increases overall salaries but disproportionately benefits men. It worsens the gender wage gap. So traditional family structure of having marriage significantly impacts the gender wage gap. But having children increases salaries equally for men and women, with no significant impact on the wage gap. So the **null hypothesis is partially rejected**: Family structure, in particular to getting married, is significantly impact the gender wage gap in management roles.

Conclusion for Hypothesis 1: _**Null:**___ Traditional family structures (partnered with children and married) do not significantly impact the gender wage gap and gender representation in

management careers. _

_**Alternative:** ___Traditional family structures (partnered with children and married) have a significant impact on the gender wage gap or gender representation in management careers._

Our results from multiple analyses support the **alternative hypothesis** that traditional family structures (married with children) significantly impact the gender wage gap and representation in management careers. From the **sectoral analysis of average salaries**, married men with children consistently earn higher salaries, particularly in the business sector, while women face significant wage gaps, even within the same family structure. Unmarried individuals (Marind=0, Chlvin=0) earn consistently lower salaries across all sectors, with unmarried women facing the steepest wage penalties. In the government sector, the gender wage gap is less pronounced, but traditional family structures still provide an advantage for both genders compared to their unmarried counterparts. The **representation analysis** shows that family structure influences gender representation, with married women with children underrepresented in high-paying roles, especially in male-dominated sectors like business. Similarly, the **control variable analysis** (e.g., employer size, working hours) indicates that larger, established organizations and high-hour workloads amplify the gender wage gap, particularly benefiting men in traditional family structures. The **Chi-Square test** confirms a significant relationship between family structure and gender representation, further validating the hypothesis. Finally, the **Regression result** confirms that traditional family structures of getting married play a significant role in the gender wage gap, suggesting the alternative hypothesis. Overall, our findings emphasize the persistent impact of family structure on gender inequality in management roles and the need for systemic interventions to address these disparities.

1.3.2 Hypothesis 2- Null: Education level does not mediate the gender wage gap in management careers for individuals who are partnered with children, in business careers.

Hypothesis 2: (Mediation Analysis)

Null: Education level does not mediate the gender wage gap in management careers for individuals in the business sector, holding marriage, children, and other factors constant.

Alternative: Education levels mediate the gender wage gap in management careers for individuals in the business sector, holding marriage, children, and other factors constant.

We are looking at the business sector here because it features performance-based pay structures, competitive environments, and less standardized salary scales compared to sectors like government or education. Also evident in the EDA from earlier, the business sector has more variability in pay across genders, and larger gaps in pay between genders within the same family arrangement. We want to see if education mediates this gap at large, for men and women in the business sector.

1.3.3 EDA For Degree Types

[ ]:
```
data = data.rename(columns={'highest_degree_type':'highest_degree_type'}) #mina
```

[ ]:
```
# Filter and prepare data
df = data[[
    'salary',
    'gender',
```

```
    'chlvin',
    'marind',
    'agegr',
    'years_at_job',
    'highest_degree_type',
    'principal_job',
    'leadership_job_titles',
    'sector',
    'emsize',
    'hrswk',
    'emrg',
    'newbus', 'earn'
]].loc[(data['sector'] == 'business')].dropna()
```

[ ]:
```
average_salaries_dgr = (
    data.copy().groupby(['principal_job', 'leadership_job_titles', 'sector',␣
 ↪'marind', 'chlvin', 'highest_degree_type'])
        .agg(
            number_of_respondents=('salary', 'size'),
            average_salary=('salary', 'mean'),
            average_earnings=('earn', 'mean')
        )
        .reset_index()
)
#mina
percent_female_dgr = (
    data.copy().groupby(['principal_job', 'leadership_job_titles', 'sector',␣
 ↪'marind', 'chlvin', 'highest_degree_type'])
        .agg(percent_female=('gender', 'mean'))
        .reset_index()
)

result_dgr = (
    average_salaries_dgr.merge(percent_female_dgr, on=['principal_job',␣
 ↪'leadership_job_titles', 'sector', 'marind', 'chlvin',␣
 ↪'highest_degree_type'])
        .assign(
            average_salary=lambda df: df['average_salary'].round(2),
            average_earnings=lambda df: df['average_earnings'].round(2),
            percent_female=lambda df: df['percent_female'].round(2)
        )
        .sort_values(['principal_job', 'sector', 'marind', 'chlvin'])
)

result_dgr[[
    'number_of_respondents',
    'principal_job',
```

```
    'leadership_job_titles',
    'highest_degree_type',
    'marind',
    'chlvin',
    'average_salary',
    'average_earnings',
    'percent_female','sector'
]].loc[(result_dgr['sector'] == 'business')].dropna() #mina
```

[ ]:
```
import altair
chart_result_dgr = altair.Chart.from_json(r"""
{
    "width": 500,
    "height": 500,
    "$schema": "https://vega.github.io/schema/vega-lite/v5.json",
    "facet": {
        "field": "leadership_job_titles",
        "title": "leadership_job_titles"
    },
    "spec": {
        "layer": [
            {
                "description": "outer data layer",
                "layer": [
                    {
                        "description": "series data layer",
                        "name": "series_a0df8f70-fe01-435f-94c8-8ab46ac8127d",
                        "layer": [
                            {
                                "description": "bar series layer",
                                "transform": [],
                                "layer": [
                                    {
                                        "description": "bar mark layer",
                                        "mark": {
                                            "type": "bar",
                                            "clip": true,
                                            "filled": true,
                                            "cursor": "pointer",
                                            "orient": "vertical"
                                        },
                                        "encoding": {
                                            "opacity": {
                                                "value": 1
                                            },
                                            "tooltip": [
                                                {
```

```
                                                    "field":␣
↪"highest_degree_type",
                                                    "type": "ordinal",
                                                    "title":␣
↪"highest_degree_type"
                                                },
                                                {
                                                    "field": "average_salary",
                                                    "type": "quantitative",
                                                    "formatType":␣
↪"NUMBER_FORMATTER",
                                                    "format": {
                                                        "format": "NUMBER",
                                                        "columnType": "NUMBER",
                                                        "numDecimalDigits": -1,
                                                        "currency": "USD",
                                                        "nanFormat": ""
                                                    },
                                                    "title": "average_salary"
                                                },
                                                {
                                                    "field": "percent_female",
                                                    "type": "quantitative",
                                                    "formatType":␣
↪"NUMBER_FORMATTER",
                                                    "format": {
                                                        "format": "NUMBER",
                                                        "columnType": "NUMBER",
                                                        "numDecimalDigits": -1,
                                                        "currency": "USD",
                                                        "nanFormat": ""
                                                    },
                                                    "title": "percent_female"
                                                }
                                            ],
                                            "color": {
                                                "field": "percent_female",
                                                "type": "quantitative",
                                                "scale": {},
                                                "legend": {
                                                    "formatType":␣
↪"NUMBER_FORMATTER",
                                                    "format": {
                                                        "format": "NUMBER",
                                                        "columnType": "NUMBER",
                                                        "numDecimalDigits": -1,
                                                        "currency": "USD",
```

```
                                                    "nanFormat": ""
                                                }
                                            },
                                            "title": "percent_female"
                                        },
                                        "x": {
                                            "field": "highest_degree_type",
                                            "type": "ordinal",
                                            "title": "highest_degree_type",
                                            "scale": {},
                                            "axis": {
                                                "grid": true,
                                                "ticks": true,
                                                "labels": true,
                                                "labelFlush": false,
                                                "labelOverlap": "greedy"
                                            }
                                        },
                                        "y": {
                                            "field": "average_salary",
                                            "type": "quantitative",
                                            "title": "average_salary",
                                            "scale": {},
                                            "axis": {
                                                "grid": true,
                                                "ticks": true,
                                                "labels": true,
                                                "labelFlush": false,
                                                "format": {
                                                    "format": "NUMBER",
                                                    "columnType": "NUMBER",
                                                    "numDecimalDigits": -1,
                                                    "currency": "USD",
                                                    "nanFormat": ""
                                                },
                                                "formatType":␣
↪"NUMBER_FORMATTER"
                                            }
                                        }
                                    }
                                }
                            ],
                            "encoding": {
                                "x": {
                                    "field": "highest_degree_type",
                                    "type": "ordinal",
                                    "title": "highest_degree_type",
```

```
                                    "scale": {},
                                    "axis": {
                                        "grid": true,
                                        "ticks": true,
                                        "labels": true,
                                        "labelFlush": false,
                                        "labelOverlap": "greedy"
                                    }
                                },
                                "y": {
                                    "field": "average_salary",
                                    "type": "quantitative",
                                    "title": "average_salary",
                                    "scale": {},
                                    "axis": {
                                        "grid": true,
                                        "ticks": true,
                                        "labels": true,
                                        "labelFlush": false,
                                        "format": {
                                            "format": "NUMBER",
                                            "columnType": "NUMBER",
                                            "numDecimalDigits": -1,
                                            "currency": "USD",
                                            "nanFormat": ""
                                        },
                                        "formatType": "NUMBER_FORMATTER"
                                    }
                                }
                            }
                        }
                    ]
                }
            ],
            "transform": []
        }
    ],
    "width": 200,
    "height": 200
  },
  "config": {
    "legend": {
        "orient": "right"
    },
    "font": "\"IBM Plex Sans\", system-ui, -apple-system,␣
↪BlinkMacSystemFont, sans-serif",
    "view": {}
```

```
    },
    "data": {
        "name": "result_dgr"
    },
    "columns": 3,
    "datasets": {
        "layer00": [
            {
                "name": "dummy",
                "value": 0
            }
        ]
    },
    "usermeta": {
        "selectionConfigs": {},
        "columnNameMappings": {
            "CHART_FOLD_KEYS": "Column",
            "CHART_FOLD_VALUES": "Values"
        }
    },
    "layer": []
}
""")
chart_result_dgr.datasets.layer00 = result_dgr.to_json(orient='records')
chart_result_dgr.display(actions=False)
```

This visualization shows the relationship between highest degree type, average salary, and the proportion of females for various leadership job titles in management roles in the business sector. Across most job categories, higher education levels (Doctorate and Professional degrees) are associated with higher average salaries, particularly for positions like C-Suite Managers, Engineering Managers, and Financial Management. The gender composition varies notably, with roles like Medical Health Managers, Personnel Managers, and Educational Administrators showing a higher proportion of females, indicated by darker blue shades. In contrast, roles such as C-Suite Managers, Computer IT Managers, and Engineering Managers have a lower proportion of females. Despite the higher salaries associated with advanced degrees, the visualization suggests that women are underrepresented in the highest-paying leadership roles. Additionally, the variation in salaries across education levels and job types highlights potential disparities that may contribute to the overall gender wage gap in management positions.

1.3.4 Weighted Least Squares and Mediation Analysis

[ ]:
```
#filter and prepare data (mina)
df = data[[
    'salary',
    'gender',
    'chlvin',
    'marind',
```

```
    'agegr',
    'years_at_job',
    'highest_degree_type',
    'principal_job',
    'leadership_job_titles',
    'sector',
    'emsize',
    'hrswk',
    'emrg',
    'newbus'
]].loc[(data['sector'] == 'business')].dropna()

#bin hours per week

columns_to_encode = [
    'agegr', 'highest_degree_type', 'principal_job',
    'leadership_job_titles', 'sector', 'emsize',
    'emrg' # If it is a categorical variable (likely)
]

# One-hot encoding for categorical variables
df_encoded = pd.get_dummies(df, columns=columns_to_encode, drop_first=True)

# Drop rows with missing values
df_encoded = df_encoded.dropna()

df_encoded.head(5)
```

[ ]:
```
df_encoded.columns
```

[ ]:
```
Index(['salary', 'gender', 'chlvin', 'marind', 'years_at_job', 'hrswk',
       'newbus', 'agegr_25', 'agegr_30', 'agegr_35', 'agegr_40', 'agegr_45',
       'agegr_50', 'agegr_55', 'agegr_60', 'agegr_65',
       'highest_degree_type_Doctorate', 'highest_degree_type_Masters',
       'highest_degree_type_Professional',
       'principal_job_management_related_occupations',
       'principal_job_non_S&E_managers',
       'leadership_job_titles_computer_it_managers',
       'leadership_job_titles_eng_managers',
       'leadership_job_titles_financial_mgmt',
       'leadership_job_titles_medical_health_manager',
       'leadership_job_titles_natural_science_manager',
       'leadership_job_titles_other_management_occupations',
       'leadership_job_titles_other_mid_level_manager',
       'leadership_job_titles_personnel_mgmt', 'emsize_100_to_499_employees',
       'emsize_10_or_fewer_employees', 'emsize_11_to_24_employees',
       'emsize_25000+_employees', 'emsize_25_to_99_employees',
```

```
       'emsize_5000_to_24999_employees', 'emsize_500_to_999_employees',
       'emrg_east_south_central', 'emrg_middle_atlantic', 'emrg_mountain',
       'emrg_new_england', 'emrg_pacific_us_territories',
       'emrg_south_atlantic', 'emrg_west_north_central',
       'emrg_west_south_central'],
      dtype='object')
```

[ ]:
```
import pandas as pd
import numpy as np
import statsmodels.api as sm
from sklearn.model_selection import train_test_split, cross_val_score
from statsmodels.stats.outliers_influence import variance_inflation_factor
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
from statsmodels.regression.linear_model import WLS
import matplotlib.pyplot as plt
```

[ ]:
```
#create an interaction term between the four highest degree types and gender␣
 ↪(mina)
df_encoded['phd_gender_interaction'] =␣
 ↪df_encoded['highest_degree_type_Doctorate'] * df_encoded['gender']
df_encoded['masters_gender_interaction'] =␣
 ↪df_encoded['highest_degree_type_Masters'] * df_encoded['gender']
df_encoded['pro_gender_interaction'] =␣
 ↪df_encoded['highest_degree_type_Professional'] * df_encoded['gender']

#log-transform the salary to stabilize variance
df_encoded['log_salary'] = np.log(df_encoded['salary'])

#step 1: check for multicollinearity using Variance Inflation Factor (VIF)
X = df_encoded.drop(['salary', 'log_salary'], axis=1)
X = sm.add_constant(X)  # Adding a constant column for VIF calculation

#calculating VIF for each feature
vif_data = pd.DataFrame()
vif_data['Variable'] = X.columns
vif_data['VIF'] = [variance_inflation_factor(X.values, i) for i in range(X.
 ↪shape[1])]

#display the VIF values
print("Variance Inflation Factors:")
print(vif_data)

#drop variables with high VIF values (>5 or >10) if necessary
#df_encoded.drop(columns=[''], inplace=True)

#split the data into training and testing sets for cross-validation
```

```
X = df_encoded.drop(['salary', 'log_salary'], axis=1)
y = df_encoded['log_salary']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,␣
 ↪random_state=42)

#build a linear regression model with Weighted Least Squares (WLS)
#create weights based on the inverse of the variance of the residuals
model_ols = sm.OLS(y_train, sm.add_constant(X_train)).fit()
residuals = model_ols.resid
weights = 1 / (residuals ** 2)

#fit the WLS model
model_wls = WLS(y_train, sm.add_constant(X_train), weights=weights).fit()

print("Weighted Least Squares Model Summary:")
print(model_wls.summary())

#evaluate model performance using cross-validation
model = LinearRegression()
cv_scores = cross_val_score(model, X, y, cv=5, scoring='neg_mean_squared_error')
mean_cv_score = -np.mean(cv_scores)
print(f"Cross-Validated Mean Squared Error: {mean_cv_score}")

#residuals vs. Fitted Values Plot
plt.scatter(model_wls.fittedvalues, model_wls.resid)
plt.xlabel('Fitted Values')
plt.ylabel('Residuals')
plt.title('Residuals vs. Fitted Values')
plt.show()

#model with Robust Standard Errors
model_robust = model_wls.get_robustcov_results(cov_type='HC3')
print("Model Summary with Robust Standard Errors:")
print(model_robust.summary())

#predict on the test set and calculate Mean Squared Error
y_pred = model_wls.predict(sm.add_constant(X_test))
test_mse = mean_squared_error(y_test, y_pred)
print(f"Test Mean Squared Error: {test_mse}")
```

[ ]:
```
/home/hexuser/.cache/pypoetry/virtualenvs/python-kernel-
OtKFaj5M-py3.9/lib/python3.9/site-
packages/statsmodels/regression/linear_model.py:1781: RuntimeWarning: divide by
zero encountered in scalar divide
  return 1 - self.ssr/self.centered_tss
/home/hexuser/.cache/pypoetry/virtualenvs/python-kernel-
OtKFaj5M-py3.9/lib/python3.9/site-
```

```
packages/statsmodels/stats/outliers_influence.py:198: RuntimeWarning: divide by
zero encountered in scalar divide
  vif = 1. / (1. - r_squared_i)
Variance Inflation Factors:
```

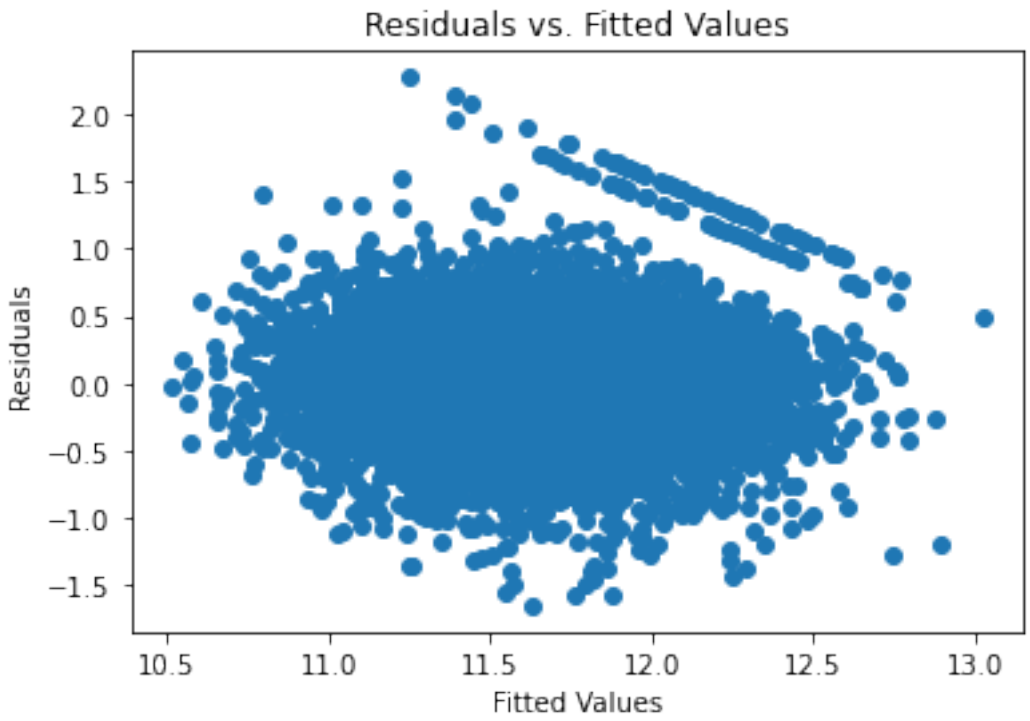
	Variable	VIF
0	const	0.000000
1	gender	1.877916
2	chlvin	1.529893
3	marind	1.417139
4	years_at_job	1.252397
5	hrswk	1.084916
6	newbus	1.062959
7	agegr_25	12.339167
8	agegr_30	19.213226
9	agegr_35	15.717276
10	agegr_40	11.673926
11	agegr_45	11.299899
12	agegr_50	10.105290
13	agegr_55	9.002383
14	agegr_60	6.314567
15	agegr_65	2.983391
16	highest_degree_type_Doctorate	1.622254
17	highest_degree_type_Masters	1.867806
18	highest_degree_type_Professional	1.918213
19	principal_job_management_related_occupations	inf
20	principal_job_non_S&E_managers	inf
21	leadership_job_titles_computer_it_managers	inf
22	leadership_job_titles_eng_managers	inf
23	leadership_job_titles_financial_mgmt	inf
24	leadership_job_titles_medical_health_manager	inf
25	leadership_job_titles_natural_science_manager	inf
26	leadership_job_titles_other_management_occupat…	inf
27	leadership_job_titles_other_mid_level_manager	1.741414
28	leadership_job_titles_personnel_mgmt	inf
29	emsize_100_to_499_employees	1.734808
30	emsize_10_or_fewer_employees	1.331727
31	emsize_11_to_24_employees	1.293426
32	emsize_25000+_employees	2.083457
33	emsize_25_to_99_employees	1.560753
34	emsize_5000_to_24999_employees	1.809846
35	emsize_500_to_999_employees	1.344849
36	emrg_east_south_central	1.171157
37	emrg_middle_atlantic	1.826072
38	emrg_mountain	1.303570
39	emrg_new_england	1.396773
40	emrg_pacific_us_territories	2.029497
41	emrg_south_atlantic	1.850755

```
42                            emrg_west_north_central   1.328737
43                            emrg_west_south_central   1.471171
44                             phd_gender_interaction   1.586905
45                          masters_gender_interaction   2.467893
46                              pro_gender_interaction   1.911060
Weighted Least Squares Model Summary:
                             WLS Regression Results
==============================================================================
Dep. Variable:             log_salary   R-squared:                       1.000
Model:                            WLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 6.822e+05
Date:                Tue, 10 Dec 2024   Prob (F-statistic):               0.00
Time:                        11:10:01   Log-Likelihood:                1903.4
No. Observations:                8435   AIC:                            -3717.
Df Residuals:                    8390   BIC:                            -3400.
Df Model:                          44
Covariance Type:            nonrobust
==============================================================================
====================================
                                                    coef    std err
t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
------------------------------------
const                                              9.0445      0.004
2443.213      0.000       9.037       9.052
gender                                            -0.1142      0.000
-230.009      0.000      -0.115      -0.113
chlvin                                             0.0185      0.000
37.794      0.000       0.018       0.019
marind                                             0.1142      0.000
241.031      0.000       0.113       0.115
years_at_job                                      -0.0006    2.3e-05
-24.715      0.000      -0.001      -0.001
hrswk                                              0.0126    2.3e-05
550.384      0.000       0.013       0.013
newbus                                             0.0323      0.001
42.503      0.000       0.031       0.034
agegr_25                                           0.1243      0.004
29.703      0.000       0.116       0.132
agegr_30                                           0.2726      0.004
65.084      0.000       0.264       0.281
agegr_35                                           0.3443      0.004
81.542      0.000       0.336       0.353
agegr_40                                           0.4223      0.004
99.929      0.000       0.414       0.431
agegr_45                                           0.4562      0.004
107.836      0.000       0.448       0.464
```

```
agegr_50                                                    0.4769      0.004
111.105      0.000       0.468       0.485
agegr_55                                                    0.5154      0.004
123.467      0.000       0.507       0.524
agegr_60                                                    0.4964      0.004
116.843      0.000       0.488       0.505
agegr_65                                                    0.4809      0.004
111.860      0.000       0.473       0.489
highest_degree_type_Doctorate                               0.2484      0.001
269.009      0.000       0.247       0.250
highest_degree_type_Masters                                 0.1221      0.001
231.646      0.000       0.121       0.123
highest_degree_type_Professional                            0.3147      0.010
31.244      0.000       0.295       0.334
principal_job_management_related_occupations                1.0241      0.001
1717.056      0.000       1.023       1.025
principal_job_non_S&E_managers                              1.7990      0.001
1976.803      0.000       1.797       1.801
leadership_job_titles_computer_it_managers                  1.6725      0.001
1752.725      0.000       1.671       1.674
leadership_job_titles_eng_managers                          1.6071      0.001
1955.166      0.000       1.606       1.609
leadership_job_titles_financial_mgmt                        0.3689      0.000
847.326      0.000       0.368       0.370
leadership_job_titles_medical_health_manager                1.4491      0.002
700.271      0.000       1.445       1.453
leadership_job_titles_natural_science_manager               1.4925      0.001
1284.605      0.000       1.490       1.495
leadership_job_titles_other_management_occupations          0.3956      0.000
1006.162      0.000       0.395       0.396
leadership_job_titles_other_mid_level_manager              -0.2794      0.001
-337.636      0.000      -0.281      -0.278
leadership_job_titles_personnel_mgmt                        0.2596      0.001
440.934      0.000       0.258       0.261
emsize_100_to_499_employees                                -0.0811      0.001
-142.953      0.000      -0.082      -0.080
emsize_10_or_fewer_employees                               -0.4120      0.002
-249.740      0.000      -0.415      -0.409
emsize_11_to_24_employees                                  -0.2634      0.001
-177.349      0.000      -0.266      -0.260
emsize_25000+_employees                                     0.0993      0.001
150.169      0.000       0.098       0.101
emsize_25_to_99_employees                                  -0.1303      0.001
-158.464      0.000      -0.132      -0.129
emsize_5000_to_24999_employees                              0.0516      0.001
80.780      0.000       0.050       0.053
emsize_500_to_999_employees                                -0.0634      0.001
```

```
-78.842      0.000      -0.065      -0.062
emrg_east_south_central                                -0.1610      0.001
-121.977      0.000      -0.164      -0.158
emrg_middle_atlantic                                    0.1489      0.001
222.984      0.000       0.148       0.150
emrg_mountain                                          -0.0316      0.001
-31.886      0.000      -0.034      -0.030
emrg_new_england                                        0.1421      0.001
204.254      0.000       0.141       0.143
emrg_pacific_us_territories                             0.2076      0.001
349.684      0.000       0.206       0.209
emrg_south_atlantic                                     0.0457      0.001
85.347      0.000       0.045       0.047
emrg_west_north_central                                -0.0233      0.001
-29.166      0.000      -0.025      -0.022
emrg_west_south_central                                 0.0276      0.001
37.925      0.000       0.026       0.029
phd_gender_interaction                                  0.0654      0.003
23.000      0.000       0.060       0.071
masters_gender_interaction                             -0.0179      0.001
-28.433      0.000      -0.019      -0.017
pro_gender_interaction                                 -0.1434      0.028
-5.140      0.000      -0.198      -0.089
==============================================================================
Omnibus:                    29255.898   Durbin-Watson:                   2.032
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1390.614
Skew:                           0.017   Prob(JB):                   1.08e-302
Kurtosis:                       1.011   Cond. No.                    1.02e+16
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The smallest eigenvalue is 1.85e-20. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
Cross-Validated Mean Squared Error: 0.1621773480219382
```

[ ]:



[ ]: Model Summary with Robust Standard Errors:

WLS Regression Results

Dep. Variable:	log_salary	R-squared:	1.000
Model:	WLS	Adj. R-squared:	1.000
Method:	Least Squares	F-statistic:	6.083e+08
Date:	Tue, 10 Dec 2024	Prob (F-statistic):	0.00
Time:	11:10:01	Log-Likelihood:	1903.4
No. Observations:	8435	AIC:	-3717.
Df Residuals:	8390	BIC:	-3400.
Df Model:	44		
Covariance Type:	HC3		

	coef	std err	t	P>|t|	[0.025	0.975]
const	9.0445	0.003	3509.588	0.000	9.039	9.050
gender	-0.1142	0.001	-200.608	0.000	-0.115	-0.113
chlvin	0.0185	0.001				

```
29.249      0.000       0.017       0.020
marind                                                  0.1142      0.001
189.158      0.000       0.113       0.115
years_at_job                                           -0.0006   3.17e-05
-17.886      0.000      -0.001      -0.001
hrswk                                                   0.0126   2.85e-05
443.091      0.000       0.013       0.013
newbus                                                  0.0323      0.001
29.647      0.000       0.030       0.034
agegr_25                                                0.1243      0.003
47.811      0.000       0.119       0.129
agegr_30                                                0.2726      0.003
104.480      0.000       0.268       0.278
agegr_35                                                0.3443      0.003
129.776      0.000       0.339       0.349
agegr_40                                                0.4223      0.003
156.071      0.000       0.417       0.428
agegr_45                                                0.4562      0.003
170.848      0.000       0.451       0.461
agegr_50                                                0.4769      0.003
172.842      0.000       0.471       0.482
agegr_55                                                0.5154      0.003
204.224      0.000       0.510       0.520
agegr_60                                                0.4964      0.004
139.181      0.000       0.489       0.503
agegr_65                                                0.4809      0.003
155.133      0.000       0.475       0.487
highest_degree_type_Doctorate                           0.2484      0.001
226.224      0.000       0.246       0.251
highest_degree_type_Masters                             0.1221      0.001
207.001      0.000       0.121       0.123
highest_degree_type_Professional                        0.3147      0.005
68.560      0.000       0.306       0.324
principal_job_management_related_occupations            1.0241      0.000
2158.290      0.000       1.023       1.025
principal_job_non_S&E_managers                          1.7990      0.001
2209.358      0.000       1.797       1.801
leadership_job_titles_computer_it_managers              1.6725      0.001
1701.572      0.000       1.671       1.674
leadership_job_titles_eng_managers                      1.6071      0.001
2196.682      0.000       1.606       1.609
leadership_job_titles_financial_mgmt                    0.3689      0.001
591.025      0.000       0.368       0.370
leadership_job_titles_medical_health_manager            1.4491      0.002
779.697      0.000       1.445       1.453
leadership_job_titles_natural_science_manager           1.4925      0.001
1603.173      0.000       1.491       1.494
```

```
leadership_job_titles_other_management_occupations     0.3956      0.000
1032.423      0.000       0.395       0.396
leadership_job_titles_other_mid_level_manager         -0.2794      0.001
-313.912      0.000      -0.281      -0.278
leadership_job_titles_personnel_mgmt                   0.2596      0.001
440.380       0.000       0.258       0.261
emsize_100_to_499_employees                           -0.0811      0.001
-97.773       0.000      -0.083      -0.080
emsize_10_or_fewer_employees                          -0.4120      0.002
-247.079      0.000      -0.415      -0.409
emsize_11_to_24_employees                             -0.2634      0.002
-147.964      0.000      -0.267      -0.260
emsize_25000+_employees                                0.0993      0.001
106.081       0.000       0.097       0.101
emsize_25_to_99_employees                             -0.1303      0.001
-100.177      0.000      -0.133      -0.128
emsize_5000_to_24999_employees                         0.0516      0.001
57.309       0.000       0.050       0.053
emsize_500_to_999_employees                           -0.0634      0.001
-62.610       0.000      -0.065      -0.061
emrg_east_south_central                               -0.1610      0.003
-48.540       0.000      -0.168      -0.155
emrg_middle_atlantic                                   0.1489      0.001
179.856       0.000       0.147       0.151
emrg_mountain                                         -0.0316      0.001
-30.515       0.000      -0.034      -0.030
emrg_new_england                                       0.1421      0.002
84.192       0.000       0.139       0.145
emrg_pacific_us_territories                            0.2076      0.001
265.590       0.000       0.206       0.209
emrg_south_atlantic                                    0.0457      0.001
72.580       0.000       0.044       0.047
emrg_west_north_central                               -0.0233      0.001
-19.776       0.000      -0.026      -0.021
emrg_west_south_central                                0.0276      0.001
33.317       0.000       0.026       0.029
phd_gender_interaction                                 0.0654      0.003
20.309       0.000       0.059       0.072
masters_gender_interaction                            -0.0179      0.001
-24.321       0.000      -0.019      -0.016
pro_gender_interaction                                -0.1434      0.029
-4.954       0.000      -0.200      -0.087
==============================================================================
Omnibus:                    29255.898   Durbin-Watson:                   2.032
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1390.614
Skew:                           0.017   Prob(JB):                   1.08e-302
Kurtosis:                       1.011   Cond. No.                    1.02e+16
```

```
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)
[2] The smallest eigenvalue is 1.85e-20. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
Test Mean Squared Error: 0.1611812719875549
/home/hexuser/.cache/pypoetry/virtualenvs/python-kernel-
OtKFaj5M-py3.9/lib/python3.9/site-packages/statsmodels/base/model.py:1888:
ValueWarning: covariance of constraints does not have full rank. The number of
constraints is 46, but rank is 45
  warnings.warn('covariance of constraints does not have full '
```

The VIF for some of the columns are very high, but we chose not to drop those columns, even though there are strong multicollinearity problems. The variables for age groups, and most of the leadership job titles, are well over 10, but dropping them would omit factors that contribute to the relationship between education level and salary among management level men and women. While multicollinearity will inflate our standard errors, it doesn't bias our coefficients, and in the context of understanding the relationships between these factors and salary, our estimates will remain valid (though with reduced precision).

In the Residual Plot from the initial WLS regression, it shows a nonlinear relationship that is not captured by our model. There is a downward trend in the residuals for higher fitted values. There is also evidence of heteroskedasticity with some of the outlying values in the residual plot, but for the most part the residuals cluster around 0.

In our WLS regression with Robust Standard Errors, we see that being female is associated with a 11.42% lower salary compared to males, holding our other factors constant. This persists even after controlling for things like education level, management job type, age group, and children. Having a doctorate is associated with a 24.84% higher salary compared to other types of degrees. Holding a master's degree is associated with a 12.21% higher salary compared to other types of degrees. Having a professional degree (like a JD, MD, or similar) is associated with a 31.47% higher salary compared to other types of degrees. The coefficient for the interaction term for having a doctorate and gender is 0.0654, which indicates that women with a PhD earn a 6.54% higher salary compared to men and women without one. Women with a master's degree have a 1.79% lower salary compared to those with a different degree. This indicates that the master's degree may not eliminate the gender wage gap, but might actually make it slightly worse. The coefficient for the interaction term between professional degrees and gender show that women with professional degrees earn a 14.34% lower salary compared to those with different degree types. This indicates that professional degrees do not reduce the gender wage gap but may make it worse. Other interesting findings from this regression are that the coefficients for age group are consistent with the idea that salaries generally increase with age, and that may contribute to higher wages among those in management roles. The coefficient for hours per week is 0.0126, which indicates that each additional hour worked per week is associated with a 1.26% increase in salary for those in management roles. Interestingly, being married is associated with an 11.42% increase in wages in business sector management roles, which could be because of social networks in these careers and household dynamics, such as the married working partner to devote more time to their career while the other partner takes on more domestic responsibilities. We also see that employer size, whether the employer is a new business,

job title and region also impact wages differentially. Certain leadership job titles, like financial management, are associated with smaller gains in salary compared to others, like natural science health managers. Working in the mountain region of the US also is associated with reductions in salary while working in the middle Atlantic is associated with gains in salary.

The WLS regression suggests that education does mediate the gender wage gap among those in management roles in the business sector, but that this effect is not consistent across all higher education types. Having a PhD can mitigate the gap, but it can also be worse for those with professional degrees, which suggests that education on its own does not fully explain the wage disparities in management careers for individuals in the business sector.

[ ]:
```
coefficients = model_wls.params #mina

# Exponentiate the coefficients
exp_coefficients = np.exp(coefficients)

# Display the exponentiated coefficients
print("Exponentiated Coefficients (interpreted as percentage change):")
print(exp_coefficients)
```

[ ]:
```
Exponentiated Coefficients (interpreted as percentage change):
const                                                 8471.728613
gender                                                   0.892115
chlvin                                                   1.018633
marind                                                   1.120953
years_at_job                                             0.999433
hrswk                                                    1.012713
newbus                                                   1.032791
agegr_25                                                 1.132341
agegr_30                                                 1.313430
agegr_35                                                 1.410933
agegr_40                                                 1.525510
agegr_45                                                 1.578072
agegr_50                                                 1.611063
agegr_55                                                 1.674381
agegr_60                                                 1.642859
agegr_65                                                 1.617594
highest_degree_type_Doctorate                            1.281970
highest_degree_type_Masters                              1.129855
highest_degree_type_Professional                         1.369896
principal_job_management_related_occupations             2.784683
principal_job_non_S&E_managers                           6.043519
leadership_job_titles_computer_it_managers               5.325726
leadership_job_titles_eng_managers                       4.988542
leadership_job_titles_financial_mgmt                     1.446153
leadership_job_titles_medical_health_manager             4.259452
leadership_job_titles_natural_science_manager            4.448362
leadership_job_titles_other_management_occupations       1.485338
```

```
leadership_job_titles_other_mid_level_manager    0.756206
leadership_job_titles_personnel_mgmt             1.296391
emsize_100_to_499_employees                      0.922069
emsize_10_or_fewer_employees                     0.662316
emsize_11_to_24_employees                        0.768467
emsize_25000+_employees                          1.104422
emsize_25_to_99_employees                        0.877800
emsize_5000_to_24999_employees                   1.052949
emsize_500_to_999_employees                      0.938575
emrg_east_south_central                          0.851269
emrg_middle_atlantic                             1.160582
emrg_mountain                                    0.968888
emrg_new_england                                 1.152711
emrg_pacific_us_territories                      1.230700
emrg_south_atlantic                              1.046765
emrg_west_north_central                          0.976963
emrg_west_south_central                          1.027965
phd_gender_interaction                           1.067546
masters_gender_interaction                       0.982236
pro_gender_interaction                           0.866380
dtype: float64
```

The exponentiated coefficients from our Weighted Least Squares regression model indicate how each variable affects the log transformed salary, and can be interpreted as percentage changes. Holding all other factors constant, the exponentiated coefficient for gender is less than 1, which indicates that being female is associated with a lower salary compared to men. The other key things to analyze with our exponentiated coefficients are the interaction terms. The interaction between gender and having a PhD is over 1, which indicates a positive relationship between PhD and salary. It is 1.072, which means that having a PhD and being female is associated with a 7.2% higher salary compared to men and women without a PhD. The interaction between having a masters and gender has an exponentiated coefficient of less than 1 at 0.985. This suggests that the salary premium for having a Master's degree is 1.45% lower for women compared to the reference group. This also indicates that the benefit of having a Master's degree is slightly reduced for women compared to men. The exponentiated coefficient for professional degrees is less than 1, and is associated with an 11.93% decrease in salary compared to the reference group.

The analysis of the exponentiated coefficients, especially those of the interaction terms, shows how gender interacts with higher degree levels to influence salary, highlighting that women with professional and master's degrees face diminished returns in management roles compared to their male counterparts, when controlling for having children in the home, years at the job, hours per week, employer size, employer region, age group, and whether it is a new business. This connects to our mediation analysis, where education level is tested as a mediator of the gender wage gap. We seek to explain how disparities in educational payoffs contribute to the overall gender wage disparity.

1.3.5 Mediation Analysis

[ ]:
```
X = df_encoded.drop(['salary', 'log_salary'], axis=1)
y = df_encoded['log_salary']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,␣
 ↪random_state=42)

X_train = X_train.reset_index(drop=True)
y_train = y_train.reset_index(drop=True)

#mina (this line from class notebook)
your_measure_of_occ_segregation_dummies = [col for col in df_encoded.columns if␣
 ↪col.startswith(('agegr_', 'principal_job_', 'leadership_job_titles_',␣
 ↪'emsize_', 'emrg_'))]

#keep key controls
controls = ['chlvin', 'years_at_job', 'hrswk','newbus', 'marind']

#drop mediators and occupational segregation variables
mediator_columns = your_measure_of_occ_segregation_dummies +␣
 ↪['highest_degree_type_Doctorate', 'highest_degree_type_Masters',␣
 ↪'highest_degree_type_Professional']
X_no_mediators = X_train.drop(columns=mediator_columns)
X_no_mediators = sm.add_constant(X_no_mediators[controls + ['gender']])

#model without mediators
model_no_mediators = sm.OLS(y_train, X_no_mediators).fit()
gender_effect_no_mediators = model_no_mediators.params['gender']

#model with mediators
X_with_mediators = sm.add_constant(X_train[controls + ['gender'] +␣
 ↪mediator_columns])
model_with_mediators = sm.OLS(y_train, X_with_mediators).fit()
gender_effect_with_mediators = model_with_mediators.params['gender']

#calculate percentage of gender effect explained by mediators
explained_percent = ((gender_effect_no_mediators -␣
 ↪gender_effect_with_mediators) / gender_effect_no_mediators) * 100

print(f"Gender effect without mediators: {gender_effect_no_mediators}")
print(f"Gender effect with mediators: {gender_effect_with_mediators}")
print(f"Percentage of gender gap explained by mediators: {explained_percent:.
 ↪2f}%")
```

[ ]:
```
Gender effect without mediators: -0.1704302903684439
Gender effect with mediators: -0.11907361025618687
Percentage of gender gap explained by mediators: 30.13%
```

The coefficient for the gender effect without mediators means that without controlling for the mediator variables (highest degree type, age group, principal job, leadership job title, employer size, and employer region), being female is associated with a 17.04% lower log-transformed salary on average, holding other factors constant.

The coefficient for the gender effect with the mediators means that when controlling for the mediator variables (highest degree types and our measures of occupational segregation), the gender effect on log salary decreases to 11.9% lower, on average. This suggests that part of the gender wage gap in the business sector for men and women in management roles can be explained by these mediators.

The percentage of the gender gap explained by the mediators is 30.13%. This means that the factors included in our mediation analysis account for about one-third of the observed gender wage-gap in management roles in the business sectors. Education level and occupational segregation significantly contribute to explaining the gender wage gap.

1.3.6 Conclusion for Hypothesis 2

_**Null:** ___Education level does not mediate the gender wage gap in management careers for individuals in the business sector, holding marriage, children, and other factors constant._

_**Alternative:** ___Education levels mediate the gender wage gap in management careers for individuals in the business sector, holding marriage, children, and other factors constant. _

The weighted least squares regression results demonstrate a statistically significant relationship between gender and higher degree type, as all the p-values for the interaction terms are less than alpha = 0.05. This offers some evidence in favor of rejecting the null hypothesis.

The results from our mediation analysis support rejecting the null hypothesis that education levels do not mediate the gender wage gap, demonstrating that while education and occupational segregation contribute to the wage disparity, they do not fully eliminate it. The remaining 69.87% of the gender wage gap suggests that other factors, such as workplace biases, motivation to work, age of children living in the home, family structure, or differences in career advancement opportunities, may also play significant roles. Therefore, while education and occupational segregation explain a substantial portion of the gap, additional factors must be explored to fully understand and address gender wage disparities in management careers.

The exponentiated coefficients show that this effect is not consistent across degree types, so while there is a significant relationship between highest degree type and salary, this effect is differential across degree types for women in particular.

For further analysis, family structure could be introduced into this analysis as well as other sector types, to see if this is consistent across all sectors, or if the return on particular degree types changes for women across sectors.

1.3.7 Implications for Research Question

Policy Recommendations:

Our data analysis for the first hypothesis shows that married individuals with children have higher salaries compared to unmarried individuals with no children. This may be due to companies

perceiving married individuals with children, particularly men, as more stable or motivated to commit to their careers, leading to higher representation, promotions, and salary increases for this group. In contrast, unmarried individuals without children often face lower salaries and fewer opportunities, as seen across all sectors. Additionally, women consistently earn and are represented less than men, a disparity that is particularly pronounced in the business and education sectors, though less so in government. To address these biases, especially in large organizations, we suggest policies must focus on reducing the influence of gender and family structure on representation rates and salary decisions. Implementing measures to monitor and correct for these biases, such as pay equity audits, transparent promotion criteria, and training programs to mitigate unconscious biases, will ensure that an employee's personal characteristics like gender and family structure play a smaller role than their skills and contributions in determining their career outcomes.

Our interaction term in our second analysis reveals women with professional degrees earn 14.34% less compared to other degree types. Private employers employing those with professional degrees (like law firms or firms hiring those with MDs) should be subject to pay audits, in thinking of ways to address this. Additionally, better policies and regulations should be put in place in the private sector to ensure equal pay for equal qualifications among men and women.

More broadly, companies should be required to report salary statistics that disclose information about salaries by gender, role, and education level. The report should also include things such as bonuses and promotions, and compare information between men and women. This would increase companies' accountability and encourage them to not create disparities. In some countries, like the UK, there is a policy like this already in place, which requires companies with over 250 employees to report salary statistics.