Национальный исследовательский университет

«Высшая школа экономики»

Московский институт электроники и математики им. А.Н. Тихонова, Москва

Направление 09.04.01. Информатика и вычислительная

техника

Бакалаврская программа "Прикладная математика"

Отчет о самостоятельной работе по

дисциплине «Методы анализа стохастических взаимосвязей»

Бригада № 9:

Пьяных Ольга Сергеевна, 3 курс, БПМ213

Москва 2024

Оглавление

1 Общая постановка задачи

1.1 Описание прикладной области и данных

Набор данных содержит исчерпывающую информацию о заработной плате в различных отраслях и регионах по всему миру, взятый с сайта Kaggle. Выборка состоит из 6685 строк, где каждая соответствует одному человеку.

№	Характеристика объекта/явления	Название переменной	Шкала измерения (одна из четырех)	Роль: целевая/объясняющая
1	Возраст (от 21 до 62)	age	относительная	объясняющая
2	Пол (Female / Male)	gender	номинальная	объясняющая
3	Уровень образования 0 - High School 1 - Bachelor Degree 2 - Master Degree 3 - Phd	education level	порядковая	объясняющая
4	Отрасль работы (191 уникальное наименование)	job title	номинальная	объясняющая

5	Количество лет опыта (от 0 до 34)	years of experience	относительная	объясняющая
6	Заработная плата (от 350 до 250.000 US)	salary	относительная	целевая
7	Страна (UK, USA, Canada, China, Australia)	country	номинальная	объясняющая
8	Национальность (10 уникальных национальностей)	race	номинальная	объясняющая
9	Работа во время учебы (0 - нет, 1 - да)	senior	номинальная	объясняющая

1.2 Основные гипотезы, которые планируется проверить в рамках исследования

Простая гипотеза

1. Зарплата зависит от количества лет опыта работы (с увеличением лет опыта работы люди становятся более квалифицированными специалистами)

Сложные гипотезы

1. Зарплата женщин с уровнем образования “Phd” растет быстрее, чем с любой другой степенью обучения (специалисты со степенью Phd всегда более востребованы, особенно, если это узконаправленная специальность, так как она указывает на высокую степень квалификации человека).
2. Зарплата у мужчин и женщин растет до определенного количества лет опыта работа, а после начинает снижаться, при этом прирост у женщин меньше, чем у мужчин (чем больше лет опыта имеет человек, тем более востребованным специалистом он является, но в тоже время увеличивается вероятность сокращения с целью переформирования штаба сотрудников, что ведет к уменьшению заработной платы. Также рассматривается гендерное неравенство).

2 Предварительный анализ собранных данных

2.1 Анализ особенностей данных: потенциальные ошибки и пропущенные значения, группы и выбросы

2.1.1 Анализ количественных переменных

Переменная	**Age**	**Years Of Experience**	**Salary**
Среднее	33.61056253740	8.0767504488	115307.175194494
Медиана	32.0	7.0	115000.0
Стандартное отклонение	7.59599442279	6.03041851629	52806.81088115
Межквартильный размах	10.0	9.0	9000.0
Верхняя квартиль	38.0	12.0	160000.0

Нижняя квартиль	28.0	3.0	70000.0
Коэффициент асимметрии	0.90356240909	0.964806399922	0.05832007334
Коэффициент эксцесса	0.1857776484	0.71166559257	-1.1666346383
Минимальное значение	21.0	0.0	350.0
Максимальное значение	62.0	34.0	250000.0
Количество наблюдений	6684	6684	6684
Количество пропущенных значения	0.0	0.0	0.0

таблица.1

- Age (возраст)
  В первую очередь заметим, что у нас нет пропущенных значений, то есть нет необходимости заполнять нулевые ячейки различными значениями. На основании анализа гистограммы, представленной на рис.1 и описательных статистик таблицы можно сделать вывод, что распределение возраста асимметрично вправо(подтверждается таблицей основных характеристик: среднее(33.6105) > медианы(32.0) и коэффициент асимметрии положительный(0.9035)). Это объясняется тем, что либо люди решили продолжить обучение, которое сказывается на недостатке времени на полноценную работу, либо полностью поменяли сферу своих интересов. Распределение имеет положительный эксцесс (0.1857), можем говорить об островершинности и наличии неоднородности.
- Years of Experience (опыт работы)
  В первую очередь заметим, что у нас нет пропущенных значений, то есть нет необходимости заполнять нулевые ячейки различными значениями. Распределение опыта работы, аналогично возрасту, ассиметрично вправо(среднее(8.0767) > медианы(7.0), коэффициент асимметрии положительный(0.9648)). Данные можно объяснить либо тенденцией сокращения людей на различных позициях для освобождения места молодым специалистам, либо выходом женщины замуж и “перекладыванием” финансовых обязанностей на супруга.
- Salary (заработная плата)
  В первую очередь заметим, что у нас нет пропущенных значений, то есть нет необходимости заполнять нулевые ячейки различными значениями. На гистограмме (рис. 1) видно, что распределение “Salary” является бимодальным, то есть существует необходимость рассматривать два отдельных кластера зарплат. Это довольно распространенное явление в данных, так как во многих странах есть две отдельные группы людей имеющих высокие и низкие доходы.

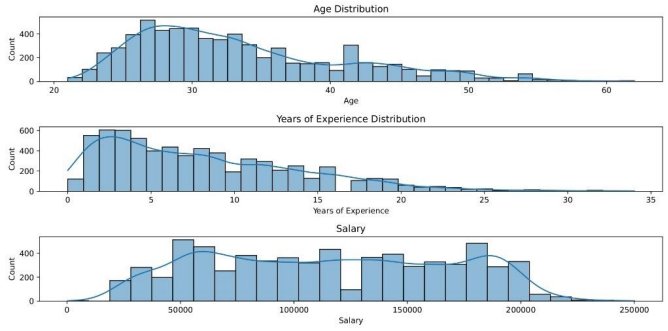


Рис 1

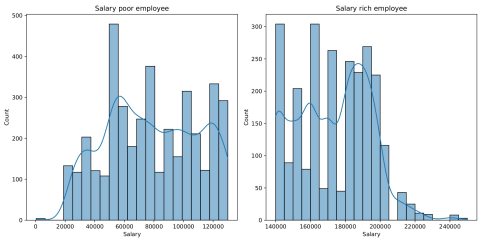


Рис. 2

Переменная	Salary_lower_part	Salary_upper_part
Среднее	78596.2942	172565.073383
Медиана	75072.0	170226.0
Стандартное отклонение	30705.89135	21085.15082
Межквартильный размах	50000.0	35000.0
Верхняя квартиль	105000.0	190000.0
Нижняя квартиль	55000.0	155000.0
Коэффициент асимметрии	0.02351	0.119082
Коэффициент эксцесса	-1.09656915	-0.5992520
Минимальное значение	350.0	140000.0
Максимальное значение	130000.0	250000.0
Количество наблюдений	4013	2521
Количество пропущенных значения	0	0

Таблица 2

Вычислим таблицу основных статистик для обоих кластеров, а также отобразим обновленные столбчатые диаграммы. Нетрудно заметить (строка в Таблице 2 “Количество наблюдений”), количество людей, относящихся к “бедному” кластеру, значительно больше, поэтому далее будем изучать его. В первую очередь, аналогично двум предыдущим пунктам, определим асимметрию распределения. Среднее(78596.2942) > медианы(75072.0), но коэффициент асимметрии близок к 0 - 0.022351, что говорит нам о незначительной асимметрии. Коэффициент эксцесса отрицательный(-1.0965), поэтому можем сказать, что распределение относительно сглаженное с закругленным пиком.

Перейдем к изучению выбросов.

```
Count of outliers for numerical values
Age outlier: 81
Years of Experience outlier: 17
Salary outlier: 0
```

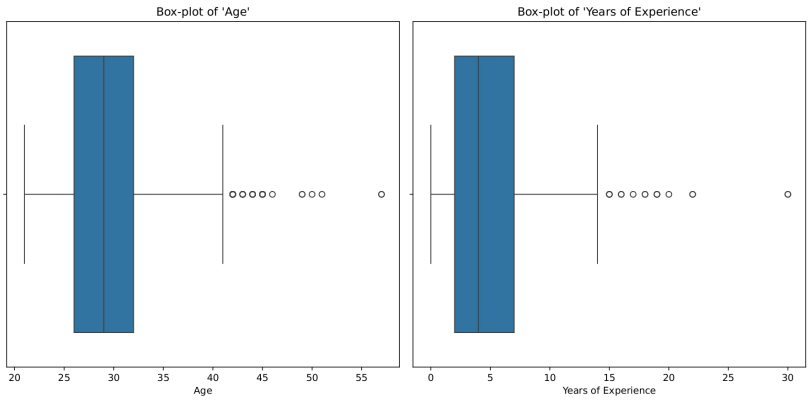
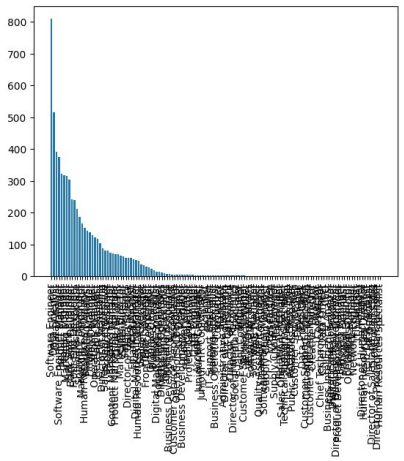


Рис 3.

Рассмотрим Рис 3. Выбросов в “Salary” не наблюдается из-за того, что в анализируемых данных представлены различные профессии с различным доходом, соответствующим действительности. Самое большое количество выбросов - в “Age”, так как средний возраст - это 30 лет, но большое количество людей работает до 60. Количество выбросов в “Years of Experience” меньше, так как некоторые люди начинают работать поздно (особенно, если получают образование уровня магистра или аспиранта) и “счет ” опыта начинается позже.

2.2.2. Анализ качественных переменных.



Column_Name	Num_Unique
Job Title	129

Рис.4

Для дальнейшего анализа проведем подготовку данных. Так как в исходной таблице больше сотни профессий (см. Рис.4), что неудобно для анализа, разобьем их на 8 групп, с которыми продолжим работу: job_client - Работа с клиентами ; job_analyze - Аналитика и данные; job_consult - Финансы и консалтинг; job_control - Управление и администрирование; job_creativity - Креативные профессии; job_marketing - Маркетинг и реклама; job_science - Исследования и разработка; job_technology - IT сфера.

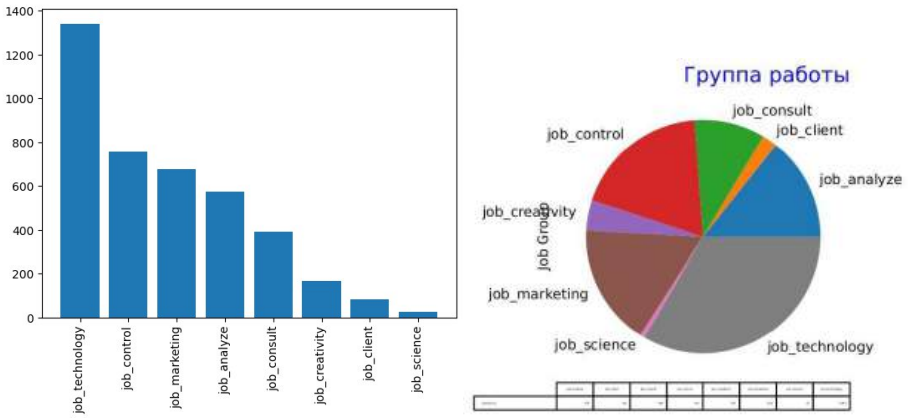


Рис. 5

На Рис.5 представлена столбчатая, а также круговая диаграммы после укрупнения данных.

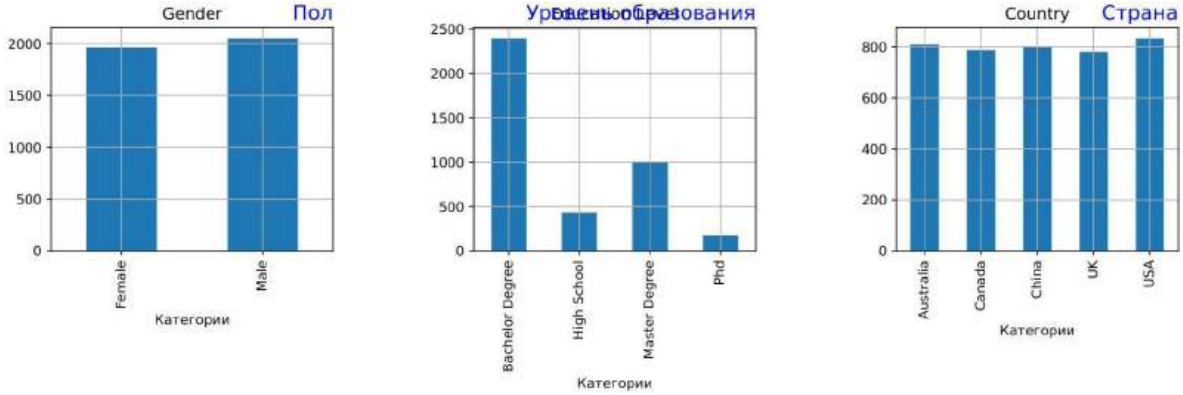


Рис 6.

На рисунке 6 представлены графики качественных переменных. Можем сразу отметить, что мужчин немного больше женщин. Такое гендерное неравенство возникает в результате того, что чаще именно женщины берут декретный отпуск по уходу за детьми и в итоге не возвращаются на работу, а семью обеспечивает мужчина.

Также стоит отметить график с уровнями образования. Подавляющее большинство людей имеют образование уровня бакалавра. Действительно, образование уровня “High School” считается недостаточным для того, чтобы быть конкурентоспособным на рынке труда, поэтому чаще всего люди идут учиться дальше. Однако после получения “Bachelor Degree” почти половина смещает свой фокус на карьеру.

Изначальные данные отличаются от исследуемых столбцом “Job Title”, так как было произведено укрупнение, в результате чего названия работ разбились на 8 подгрупп, в большинстве которых не требуется образование выше “Bachelor Degree”. Отсюда следует вывод о “непопулярности“ получения степеней “Master Degree” и “Phd”. На последней гистограмме представлено количественное распределение людей между 5 странами, которые попали в изучаемые данные. Столбцы достаточно похожи между собой, что говорит о высоком качестве данных.

Рис 7.

Опираясь на Рис 7, проанализируем информацию о “Race”. Большинство людей являются белыми или азиатами. Так как в данных представлены всего 5 стран, где превалируют именно эти расы, такое распределение не вызывает вопросов. Также стоит отметить, что все страны, участвующие в опросе, являются развитыми, а значит там проживает достаточно большое количество мигрантов, что объясняет присутствие других рас в данных.

Проведем анализ переменной “Senior”. Нетрудно заметить, что подавляющее большинство человек в данном датасете не работали пока учились в школе. Это говорит нам о том, что в приоритете было хорошее образование. В отличие от 1000 людей, которые, по семейным обстоятельствам, например, небогатая семья или отсутствие одного или обоих родителей, или другим личным причинам начали зарабатывать свои первые деньги еще в школе. Рассмотрим распределение участников исследования по профессиям. Большинство опрошенных занимаются распространенными профессиями, которые, как считается, способны обеспечить приличный уровень жизни. Меньшее количество людей выбирают творческие специальности, науку и работу с людьми. Именно эти профессии, как известно, несут в себе больше рисков, стресса и не гарантируют высокий заработок.

2.2 Анализ статистической связи.

2.2.1 Графический анализ пары «целевая переменная – качественная объясняющая переменная».

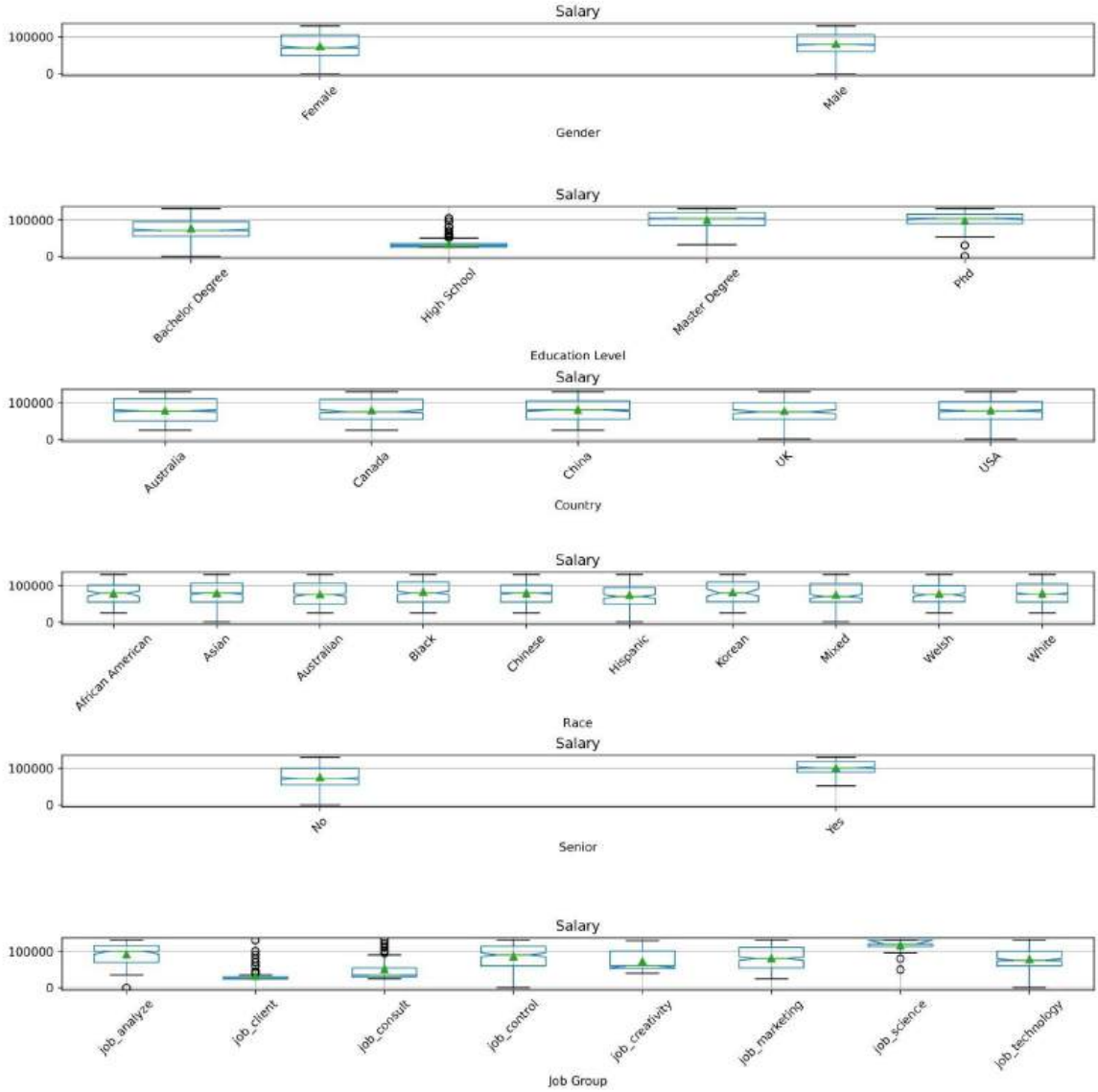


Рис. 8. Категоризированные диаграммы Бокса-Вискера.

```
Критерий Крускала-Уоллиса для переменных 'Salary' и 'Gender'
KruskalResult(statistic=37.1297001398008, pvalue=1.1052738714282362e-09)
Критерий Крускала-Уоллиса для переменных 'Salary' и 'Education Level'
KruskalResult(statistic=1439.186434548148, pvalue=9.245631134223e-312)
Критерий Крускала-Уоллиса для переменных 'Salary' и 'Country'
KruskalResult(statistic=3.754007694847565, pvalue=0.4403196116934148)
Критерий Крускала-Уоллиса для переменных 'Salary' и 'Race'
KruskalResult(statistic=16.594918873673524, pvalue=0.055450533579641685)
Критерий Крускала-Уоллиса для переменных 'Salary' и 'Senior'
KruskalResult(statistic=231.15428086001202, pvalue=3.3389956973124232e-52)
Критерий Крускала-Уоллиса для переменных 'Salary' и 'Job Group'
KruskalResult(statistic=650.9893129678211, pvalue=2.526902196750044e-136)
```

Рис. 9 Критерий Крускала-Уоллиса для переменных

Рассмотрим графики рис. 8 – диаграммы отражают выбросы и средние значения. По первой диаграмме можно сделать вывод, что “Salary” сильно зависит от “Geder”, причем причем

среднее значение у мужчин выше. Наличие этой взаимосвязи также подтверждает значение теста Крускала-Уоллиса (pvalue = 1.10527 * 10^(-9) < 0.05). Такое поведение можно объяснить культурным различием и разницей в ожидаемом поведении для мужчин и женщин в обществе - часто “добытчиком” в семье считается именно мужчина.

Рассмотрим диаграммы “Country”. Среднее значение зарплаты не зависит от стран, в которых проживает или работает человек. В подтверждение можно привести критерий теста Крускала-Уоллиса (pvalue=0.5545 > 0.05), что говорит нам о невозможности отвержения нулевой гипотезы об отсутствии влияния качественной переменной на целевую. Это происходит в силу рассмотрения стран, находящихся на одном уровне развития.

Для диаграммы “Race” среднее значение заработной платы не зависит от названия расы, что подтверждается тестом Крускала-Уоллиса (pvalue=0.4403 > 0.05), в результате которого невозможно отвергнуть нулевую гипотезу об отсутствии связи между качественной и целевой переменными. Объясняется это доступностью различных профессией для любых рас в изучаемых странах.

Далее рассмотрим диаграмму “Education Level”. Выбросы в случае образования уровня “High School” обусловлены тем, что либо человек имеет талант, что позволяет ему получать более высокие должности без высшего образования, либо родственниками являются директора компаний, которые могут помочь с трудоустройством на высокую должность с минимальным опытом работы. В случае уровня образования “Phd” человек либо является недавним выпускником с небольшим опытом работы, либо не работает в сфере, по которой получил образование. Анализируя диаграмму Бокса-Вискера в целом, заметим, что средние значения у каждой категории различны. Наличие данной взаимосвязи подтверждается тестом Крускала-Уоллиса (pvalue = 9.2456 * 10^(-312) < 0.05).

2.2.2 Графический анализ пары «числовая зависимая переменная – числовая независимая переменная».

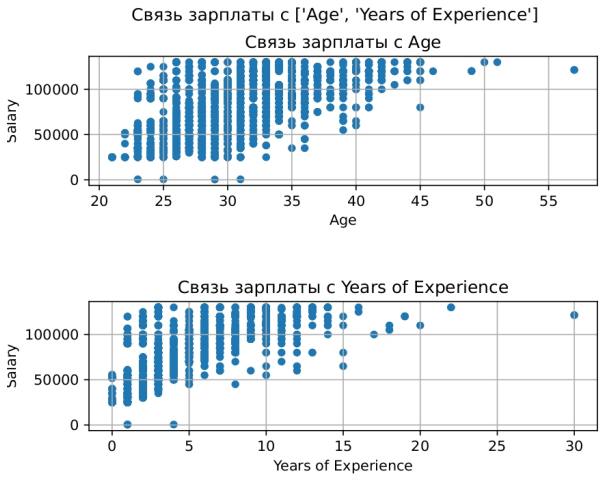


Рис 10

При первом рассмотрении графиков можно заметить, что они достаточно похожи похожи между собой, то есть можно предположить, что перед нами

мультиколлинеарные переменные.

переменная	**Age-YoE**
корреляция Пирсона	0.914837
p-value	0.000000
корреляция Спирмена	0.891149
p-value корреляция Спирмена	0.000000
корреляция Кендалла тау	0.779776
p-value корреляция Кендалла тау	0.0000

Таблица 3

В подтверждение рассуждениям выше приведем таблицу, где показаны коэффициенты корреляция Пирсона, Спирмена и Кенделла тау, а также значения pvalue. Значения корреляций достаточно большие, в тоже время pvalue < 0.05 для каждого теста, что говорит нам о связи между возрастом и количеством лет опыта работы.

Рассмотрим каждую диаграмму рассеивания по отдельности. На первой представлена явная зависимость зарплаты от возраста. Действительно, обращаясь к предметной области, можно отметить, что с возрастом люди получаются больше опыта. Аналогично для второй диаграммы. Более востребованными специалистами являются люди, у которых много лет опыта за плечами.

переменная	**Age**	**Years of Experience**
корреляция Пирсона	0.591528	0.745051
p-value	0.000000	0.000000
корреляция Спирмена	0.595873	0.783892
p-value корреляция Спирмена	0.000000	0.000000
корреляция Кендалла тау	0.448430	0.640975

p-value корреляция Кендалла тау	0.0000	0.000000

Таблица 4

Для формальной проверки гипотез воспользуемся таблицей 4, где представлены коэффициенты корреляции Пирсона и Спирмена, а также Кендалла тау для связи “Salary”-”Age” и “Salary”-”Years of Experience”. Заметим, что в обоих случаях коэффициенты различных корреляция достаточно велики, а значения pvalue наоборот стремятся к нулю. Тогда мы можем утверждать, что обе переменные являются сильно значимыми. Положительные знаки также верны в силу описанным выше рассуждениям.

2.2.3 Анализ статистической взаимосвязи между независимыми переменными**.**

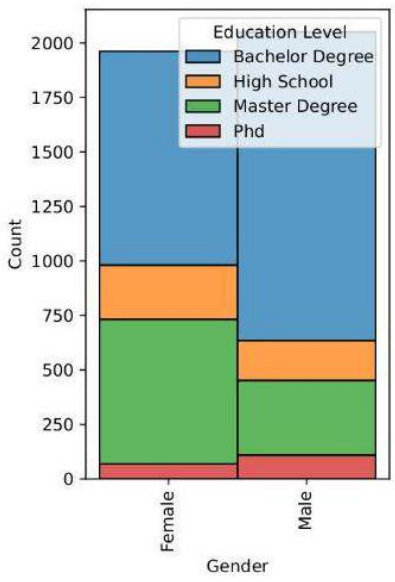


рис 11a

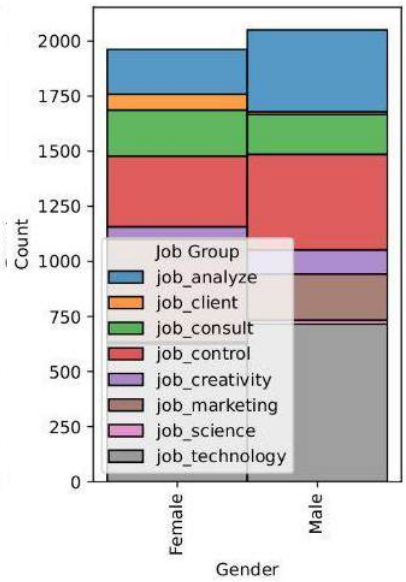


рис 11b

Рассмотрим гистограмму, связывающую гендер и уровень образования (рис. 11a).. Заметим, что “Bachelor Degree” чаще получают мужчины, нежели женщины. Аналогичная ситуация для “Phd”. Но с другой стороны женщины чаще после получения степени бакалавра продолжают обучение и становятся магистрами. Действительно, из-за гендерного неравенства женщинам труднее найти работу, но с получением степени “Master Degree” разрыв сокращается.

На диаграмме (рис 11б) показана связь между гендером и группой профессии, в которой работает человек. Аналогично первому графику количество мужчин в целом преобладает, но количество женщин лидирует в таких группах как: client, creativity, consult, marketing. Это объясняется тем, что в данных группах превалирует общение с клиентами, а также творческие подходы, что не совсем свойственно “мужским” профессиям.

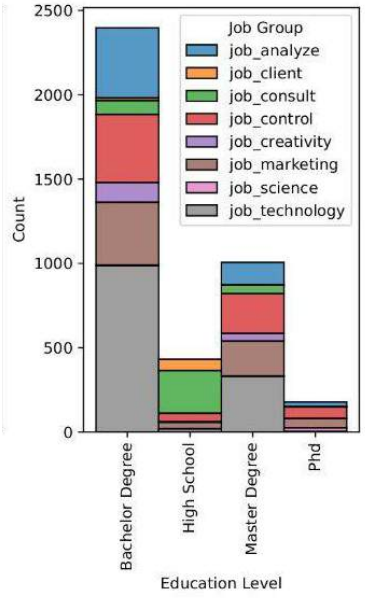


Рис 12a

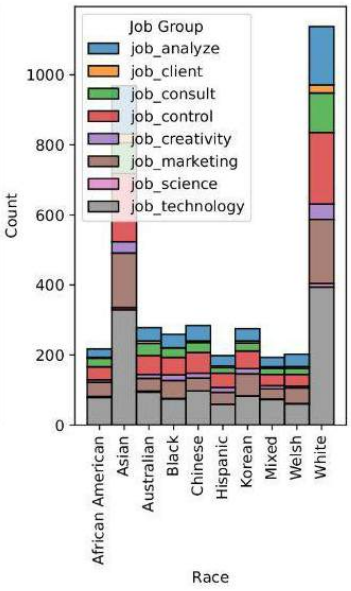


Рис 12b

Обратим внимание на связь образования и расы с профессией (рис. 12a и 12b). Профессии действительно требуют разный уровень образования и это отражено на графике. Люди без высшего образования ограничены в выборе, поэтому чаще всего выбираю такие сферы как: consult, client, marketing. Сфера control достаточно необычна для данного вида образования, но может быть объяснена либо родственниками, которые помогли с трудоустройством на высокую должность, либо выдающимися талантами. Специалисты с высшим образованием предпочитают сферу control и marketing, где сильным преимуществом является образование не ниже уровня “Master Degree”, так как происходит существенный прирост к заработной плате. У степеней “Bachelor Degree” и “Master Degree” выделяется сфера technology, так как на данный момент она очень востребована из-за развития различных технологий, и ощущается острая нехватка специалистов. В отличие от уровня образования раса не так сильно влияет на выбор сферы деятельности. Исключением являются asian и white, так как это доминирующие расы. Но также как и с уровнем образования, бывшие студенты разных рас стремятся в сферу technology.

Для анализа силы связи между качественными переменными, мы:

- построим таблицу кросс-табуляции;
- приведем значения статистики хи-квадрат;
- приведем значения статистики V-Крамера.

- Связь "Gender - Education Level":

Статистика Xu^2	197.927307
Xu^2 test p-value	0.000000
V Крамера	0.157037

	Bachelor	**High School**	**Master**	**PhD**	**All**

	Degree		Degree		
Female	981	249	663	69	1962
Male	1416	183	343	109	2051
All	2397	432	1006	178	4013

Левая таблицы: кросс-табуляция, правая таблица: статистики

Таблицы 5

Нулевая гипотеза об отсутствии взаимосвязи переменных отвергается (p-value = 0.000 < 0.05). Посмотрим на значение V Крамера, оно равно 0.1570, то есть, несмотря на то, что результат статистически значим, связь между переменными достаточно слабая.

- Связь "Gender - Job Group":

	job analize	job client	job consult	job control	job creativity	job marketing	job science	job technology	all
Female	203	72	209	321	57	466	9	625	1962
Male	372	12	181	434	110	209	18	715	2051
All	575	84	390	755	167	675	27	1340	4013

Статистика Xu^2	233.307712
Xu^2 test p-value	0.00000
V Крамера	0.170496

Левая таблицы: кросс-табуляция, правая таблица: статистики

Таблицы 6

Нулевая гипотеза об отсутствии взаимосвязи переменных отвергается (p-value = 0.000 < 0.05). Посмотрим на значение V Крамера, оно равно 0.1704, то есть, несмотря на то, что результат статистически значим, связь между переменными достаточно слабая.

- Связь "Education Level - Job Group":

Статистика Xu^2	2354.894563
Xu^2 test p-value	0.000000
V Крамера	0.383020

	job analize	job client	job consult	job control	job creativity	job marketing	job science	job technology	all

bachelor degree	416	16	82	403	117	373	4	986	2397
high school	0	67	253	50	4	38	0	20	432
master degree	132	1	52	235	46	208	2	330	1006
PhD	27	0	3	67	0	56	21	4	178
All	575	84	390	755	167	675	27	1340	4013

Левая таблицы: кросс-табуляция, правая таблица: статистики

Таблицы 7

Нулевая гипотеза об отсутствии взаимосвязи переменных отвергается (p-value = 0.000 < 0.05). Посмотрим на значение V Крамера, оно равно 0.3820. Результат является умеренным, связь между полями средней силы.

- Связь "Race - Job Group":

Job Group	job_analyze	job_client	job_consult	job_control	job_creativity	job_marketing	job_science	job_technology	All
Race\Job Group									
African American	24	3	24	37	7	41	3	78	217
Asian	138	25	88	195	32	156	6	329	969
Australian	38	7	35	54	11	36	3	94	278
Black	38	1	27	50	16	51	2	74	259
Chinese	44	5	28	59	14	37	0	97	284
Hispanic	29	5	16	41	14	34	0	59	198
Korean	35	6	23	50	15	63	0	83	275
Mixed	27	4	18	32	9	29	2	72	193
Welsh	35	5	18	33	5	45	1	60	202
White	167	23	113	204	44	183	10	394	1138
All	575	84	390	755	167	675	27	1340	4013

Связь	Race - Job Group
Статистика $\chi 2$	59.977818
$\chi 2$ test p-value	0.953953
V Крамера	0.043223

Левая таблицы: кросс-табуляция, правая таблица: статистики

Таблицы 8

Нулевая гипотеза об отсутствии взаимосвязи переменных не отвергается (p-value = 0.9539 > 0.05). Значение V Крамера равно 0.0432, что подтверждает наличие очень слабой между переменными.

2.2.4 Предварительная проверка гипотез

1. Для первой гипотезы все достаточно очевидно. Из вышеприведенного анализа мы можем утверждать, что связь между возрастом и заработной платой действительно присутствует. Самым наглядным примером является диаграмма рассеивания (рис 10), а также коэффициенты корреляции, представленные в таблице 4
2. Для данной гипотезы мы может установить, что связь между гендером и уровнем образования существует (таблица 6)), но значение V Крамера было достаточно мало, что говорит нам о необходимости дальнейшего исследования.
3. Связь между гендером и заработной платой подтверждается тестом Крускала-Уоллиса, а также диаграммой Бокса-Вискера (рис 8-9). Определить пороговое значение без исследования модели не представляется возможным, но можно утверждать, что в заработной плате присутствует некоторый пик, после которого происходит падение (рис 1)

3 Проверка гипотез с помощью моделирования

3.1. Построение базовой модели.

В качестве базовой модели используется регрессионная модель с линейным включением всех значимых переменных, за исключением страны и расы, так как нет значимой корреляционной связи с зависимой переменной. Это может быть подтверждено с помощью диаграммы Бокса-Вискера, а также теста Крускала-Уоллиса (рисунок 9).

Выборка случайным образом разделяется на две части: обучающую (80% общего объёма данных) и тестовую (20%), что позволит проверить прогностические свойства модели. Для создания базовой модели мы также преобразуем некоторые порядковые и номинальные в фиктивные (dummy) переменные, которые будут служить индикаторами.

Тогда уравнение регрессии можно записать следующим способом: sa;ary=

$$a_0 + a_1"years\ of\ experience" + a_2"Gender\ female" + a_3"job\_analyze" + a_4"job\_control" +$$

$$a_5"job\_consult" + a_6"job\_creativity" + a_7"job\_marketing" + a_8"job\_science" +$$

$$+ a_9"job\_technology" + a_{10}"Senior\_Yes" + a_{11}"Education\ Level\_Bachelor\ Degree" +$$

$$a_{12}"Education\ Level\_Master\ Degree" + a_{13}"Education\ Level\_Phd"$$

Базовый вариант: мужчина, работающий в группе control job, без высшего образования (образование “High School”), не на позиции senior. Дальнейшие результаты будем сравнивать именно с этой моделью. С помощью метода fit производим оценку базовой модели:

```
========================================
R-squared:                        0.742
Adj. R-squared:                   0.741
F-statistic:                      1093.
Prob (F-statistic):               0.00
Log-Likelihood:                 -62142.
AIC:                          1.243e+05
BIC:                          1.244e+05
```

Рис 13

	vars	VIF
0	const	239,7857974
1	Gender_Female	1,117026769
2	Education Level_Bachelor Degree	5,412186913
3	Education Level_Master Degree	5,250896796
4	Education Level_Phd	2,05144946
5	Job Group_job_analyze	8,80494807
6	Job Group_job_control	10,63282581
7	Job Group_job_consult	5,502360634
8	Job Group_job_creativity	3,729625146
9	Job Group_job_marketing	9,451387146
10	Job Group_job_science	1,578058766
11	Job Group_job_technology	14,84920328
12	Senior_Yes	1,264565364
13	Age	6,498942438
14	Years of Experience	7,369632064

Рис 14

F-statistic- проверка гипотезы, что все коэффициенты линейной регрессионной кроме константы равны 0 . Значимости F статистики близка к 0, значит мы можем отвергнуть нулевую гипотезу и модель имеет право на существование.

Проверяем модель на мультиколлинеарность. Необходимо это сделать, так как, если упустить мультиколлинеарность, дисперсии оценок могут сильно вырасти, что повлияет на точность анализа.

Рассмотрим представленную выше таблицу, на которой изображены переменные и их параметры VIF. Сразу можно выделить несколько претендентов на удаление: Job Group_job_consult, Job Group_job_creativity, Job

Group_job_maerketing, Job Group_job_technology, Job Group_job_control, Job Group_job_analyze, Age, Years of Experience, Education Level_Bachelor Degree и Education Level_Master Degree. Удаление по одной из сфер работ не принесет нам хороших результатов, так как изучаемые данные будут неполными, аналогично для двух видов образования. Большой параметр VIF в сферах работ мог возникнуть из-за разделения данных по сферам, так как некоторые профессии стоят на периферии 2 или более различных групп, но в нашем анализе необходимо было определить их только в одну.

Таким образом, мы рассмотрим только поочередное удаление переменных Age и Years of Experience и пересчитаем для полученных моделей новые параметры. Наилучшую модель будем выбирать опираясь на AIC оценку.

	vars	VIF
0	const	76,22018252
1	Gender_Female	1,114365051
2	Education Level_Bachelor Degree	5,220872587
3	Education Level_Master Degree	4,834933707
4	Education Level_Phd	1,929516907
5	Job Group_job_analyze	1,621586267
6	Job Group_job_client	1,291789929
7	Job Group_job_consult	1,984091215
8	Job Group_job_creativity	1,298703611
9	Job Group_job_marketing	1,591815465
10	Job Group_job_science	1,149626191
11	Job Group_job_technology	2,051626738
12	Senior_Yes	1,247701045
13	Age	1,444427871

	vars	VIF
0	const	24,00747669
1	Gender_Female	1,116944585
2	Education Level_Bachelor Degree	5,394700679
3	Education Level_Master Degree	5,19962276
4	Education Level_Phd	2,01397408
5	Job Group_job_analyze	1,596949887
6	Job Group_job_client	1,2890632
7	Job Group_job_consult	1,984548767
8	Job Group_job_creativity	1,291120727
9	Job Group_job_marketing	1,59217974
10	Job Group_job_science	1,150546593
11	Job Group_job_technology	2,03571466
12	Senior_Yes	1,264392271
13	Years of Experience	1,63794372

Рис 15

Заметим, что после удаления по одной переменной в моделях не осталось мультиколлинеарных связей, что говорит нам о том, что мы переменные для удаления выбраны верно.

R-squared:	0.562
Adj. R-squared:	0.560
F-statistic:	315.4
Prob (F-statistic):	0.00
Log-Likelihood:	-36404.
AIC:	7.284e+04
BIC:	7.292e+04

R-squared:	0.687
Adj. R-squared:	0.686
F-statistic:	540.2
Prob (F-statistic):	0.00
Log-Likelihood:	-35864.
AIC:	7.176e+04
BIC:	7.184e+04

Рис 16 Левая таблица - базовая модель без “Years of Experience”, правая таблица - без “Age”

Проведем сравнение параметров R^2 и AIC. Значения таблицы, которая не содержит столбца “Age”, лучше (AIC: 7.184e4 <7.284e4; R^2: 0.687>0.560). При дальнейшем анализе будем использовать модель без “Age”.

Проведем проверку базовой модели на наличие гетероскедастичности. Результатом проверки тестом Уайта стало утверждение об отвержении нулевой гипотезы о гомоскедастичности модели (из таблицы видно, что значение LM_P < 0.1).

3.2. Проверка гипотез с помощью моделирования

Результаты проверки первой гипотезы: не меняя базовую модель, проверяем коэффициент при параметре “Years of Experience”. Он равен 6243.98, что говорит нам о том, что мы отвергаем нулевую гипотезу о независимости зарплаты от количества лет опыта работы. Значит, чем больше опыта у человека, тем выше его заработная плата. Гипотеза подтвердилась.

	0
LM	1429,559575
LM_P	0
F	40,75490319
F_P	0

Рис 17

	coef
const	3.076e+04
Gender_Female	-4398.5368
Education Level_Bachelor Degree	1.689e+04
Education Level_Master Degree	2.69e+04
Education Level_Phd	2.357e+04
Job Group_job_analyze	1.771e+04
Job Group_job_client	-6926.5418
Job Group_job_consult	-3581.4514
Job Group_job_creativity	1940.1101
Job Group_job_marketing	-3348.3615
Job Group_job_science	5903.0761
Job Group_job_technology	6601.0789
Senior Yes	-2739.2598
Years of Experience	6243.9800

Рис 18

Для гипотезы 2 введем переменные: *Phd_F* и *Not_Phd_F*. Тогда уравнение регрессии примет вид:

$$Salary = a_0 + a_1 "years\ of\ experience" + a_2 "Gender\ female" + a_3 "job\_analyze" + a_4 "job\_client" +$$

$$a_5 "job\_consult" + a_6 "job\_creativity" + a_7 "job\_marketing" + a_8 "job\_science" +$$

$$+ a_9 "job\_technology" + a_{10} "Senior\_Yes" + a_{11} "Education\ Level\_Bachelor\ Degree" +$$

$$a_{12} "Education\ Level\_Master\ Degree" + a_{13} \text{"Education Level\_Phd"} + a_{231} "Phd\_F" +$$
$$a_{232} "Not\_Phd\_F"$$

Если вторая гипотеза верна, то $a_{231} > a_{232}$, $a_{231} > 0$.

Результаты проверки второй гипотезы: рассмотрим изображение 19. По итогам моделирования, $a_{231} > a_{232}$, а значит гипотеза верная. Такие результаты можно объяснить тем, что после окончания обучения именно девушки с PhD являются лучшими специалистами. Также они старше и, как правило, серьезнее относятся к работе и выполнению обязанностей.

	coef
const	3.083e+04
Gender_Female	-3640.3336
Education Level_Bachelor Degree	1.689e+04
Education Level_Master Degree	2.694e+04
Education Level_Phd	2.275e+04
Job Group_job_analyze	1.77e+04
Job Group_job_client	-6915.0204
Job Group_job_consult	-3601.9003
Job Group_job_creativity	1915.9718
Job Group_job_marketing	-3341.2628
Job Group_job_science	6276.8434
Job Group_job_technology	6582.7381
Senior_Yes	-2746.3852
Years of Experience	6238.3475
Phd_F	2794.0612
Not_Phd_F	846.2724

Рис 19

	coef
const	4.608e+04
Gender_Female	-3679.1415
Education Level_Bachelor Degree	1.716e+04
Education Level_Master Degree	2.617e+04
Education Level_Phd	2.364e+04
Job Group_job_analyze	1.919e+04
Job Group_job_client	-7638.0639
Job Group_job_consult	-3061.7679
Job Group_job_creativity	3660.3750
Job Group_job_marketing	-1300.9370
Job Group_job_science	7489.8889
Job Group_job_technology	8224.1580
Senior_Yes	-3602.3744
Years of Experience	4588.8819
Female_YoE	1.262e+04
Male_YoE	1.359e+04

Рис 20

коэффициенты моделей

Составим для третьей гипотезы новую формулу, включающую в себя две новые переменные : "*Female_YoE*" и "*Male_YoE*", где Years of Experience подбирается экспериментально.

$Salary = a_0 + a_1$"years of experience" + a_2"Gender female" + a_3"job_analyze" + a_4"job_client" + a_5"job_consult" + a_6"job_creativity" + a_7"job_marketing" + a_8"job_science" ++ a_9"job_technology" +a_{10}"Senior_Yes" + a_{11}"Education Level_Bachelor Degree" +a_{12}"Education Level_Master Degree" + a_{13}"Education Level_Phd" +a_{232}"*Female_YoE*"+a_{231}"*Male_YoE*". Если вторая гипотеза верна, то $a_{231} > a_{232}$, $a_{231} > 0$.

Результаты проверки третьей гипотезы: рассмотрим изображение 20. По итогам моделирования коэффициенты при новых переменных больше 0, что говорит о росте з\п с ростом количества опыта. Также заметим, что коэффициент при переменной, соответствующей мужчинам, больше, чем коэффициент соответствующий женщинам. Гипотеза верна. В начале карьеры работники быстро развиваются и набирают опыт и в итоге заслуживают повышения и заработная плата значительно растет. После выхода на комфортный уровень з\п, позволяющий содержать семью, скорость роста постепенно падает, а ближе к пенсии и вовсе снижается- работники берут меньше часов и дополнительных задач, не так сильно мотивированы. В то же время из-за гендерного неравенства у мужчин з\п растет быстрее.

Теперь запишем итоговую модель, которая включает в себя переменные из обеих сложных гипотез. Это итоговая модель:

$Salary = a_0 + a_1$"years of experience" + a_2"Gender female" + a_3"job_analyze" + a_4"job_client" + a_5"job_consult" + a_6"job_creativity" + a_7"job_marketing" + a_8"job_science" ++ a_9"job_technology" +a_{10}"Senior_Yes" + a_{11}"Education Level_Bachelor Degree" +a_{12}"Education Level_Master Degree" + a_{13}"Education Level_Phd" +a_{1-1}"yearsOfExperience – PHD"+a_{1-2}"yearsOfExperience – Bachelor" a_{2-1}" "*Female_YoE*"+a_{2-2}" "*Male_YoE*"

3.3. Оптимизация итоговой модели, сравнение качества моделей.

Номер\Критерий	R^2	Adj R^2	Akaike
Базовая модель	0.687	0.686	7.176e+04
Первая модифицированная	0.687	0.686	7.176e+04
Вторая модифицированная	0.700	0.699	7.162e+04
Итог	0.700	0.699	7.162e+04

Итоговая модель прошла оптимизации за счет пошагового удаления незначимых переменных(Not_PhD_F) . Для финального варианта оценивается качество модели с использованием критерия Akaike и adjusted R-sq. Итоговая модель оказалось самой лучшей(ее R^2 самый большой из рассматриваемых, а akaike наоборот самый маленький.

3.4. Проверка прогностических способностей модели

Посчитали значения прогнозов для элементов тестовой выборки и построить для них центральные доверительные интервалы на основе нормального распределения для доверительной вероятности 95%. Итоговая - до оптимизация, финальная- после

Полученные результаты говорят о том, что с точки зрения доверительной вероятности и погрешностей лучшая модель - базовая. Доверительная вероятность очень высока - более 95%, а погрешность меньше чем у промежуточной 1

Номер\Критерий	Среднеквадратическая погрешность	Абсолютная погрешность	Доверительная вероятность
Базовая	585.5677	9897119.164	0,9651307597
Промежуточная 1	587,9774318	9961169,014	0,9651307597
Промежуточная 2	921,7627636	17617038,11	0,7907845579
Итоговая	613,1067029	10970069,3	0,9613947696
Финальная	607,0784651	10769852,03	0,9626400996

Таблица 4. Сравнение прогностических способностей моделей

4. Заключение

По итогам проверки все три гипотезы подтвердились. Итоговая гипотеза, включающая в себя переменные ‘Female_YoE', 'Phd_F', ‘Male_YoE показала себя хуже базовой и ее прогнозы не такие точные.