GPG

https://github.com/meGregV/GPG

2021-01-06

INTRODUCTION

Men and women doing the same work at the same company must be equally compensated by law. The gender pay gap (GPG), defined as the difference in mean or median pay between men and women, has received growing attention from researchers, policy makers and the wider public. GPG is a complex aggregate measure resulting from differences in labor participation, education, skills, personal preferences, social and cultural norms, access and opportunities, responsibilities, levels of risk aversion and forward planning plus other characteristics such as age, ethnicity, etc.

While the pay gap has been falling in the last two decades, it remains substantial. At the same time, it is evident that the gender pay gap is not necessarily driven by employers discrimination as numerous other explanations have been proposed in the literature. Understanding the causes of the gender pay gap is crucial to provide adequate policy recommendations to reduce the gap. Misunderstanding gender pay gap may put unnecessary burden on companies resulting in ill-advised measures to close the gap. In reality it might do little to address the issue of gender discrimination akin to trying to reduce lung cancer while ignoring critical risks factors such as cigarette smoking or asbestos exposure.

The literature on GPG has been vibrant and many causes have been researched. Below are some examples:

- Differences in psychological attributes, such as attitudes towards risk, competition and negotiation.
- Family and fertility decisions as having children typically leads to career interruption for women, but not men.
- Still prevailing differences in education, while better educational access for women has narrowed the gender educational gap in recent decade.
- The difficulty for women to combine work and family, especially to work long or particular hours makes women favor jobs with greater flexibility, fewer hours, and closer proximity to home ("commute bias").

While some of these factors could be driven inherently by gender discrimination as well, the research that tried to control as many of these factors as possible has found that the GPG does level off. Therefore, if GPG is partly caused by women's willingness to accept lower wages in exchange for greater flexibility and shorter commutes, then reducing the gap should focus on the reasons for women favoring more flexible jobs.

Equipped with a publicly available UK employers data set, this project will have another look at GPG and construct a prediction model. As a stand along metric, modeled GPG could still be beneficial in highlighting inherent trends in gender discrimination, especially when comparing to the peer companies.

DATA

The primary dataset is sourced from UK Goverment Equalities office that helps to identify and manage gender pay gap by collecting various pay metrics by gender from UK employers. This is mandatory process

for UK organizations with over 250 employees and optional for those with smaller headcount. The following metrics in percent are collected:

$$mean(median)\ GPG\ in\ hourly\ pay : \mu_{GPG} = \frac{\mu_{hourly\_pay\_male} - \mu_{hourly\_pay\_female}}{\mu_{hourly\_pay\_male}} * 100$$

$$mean(median)\ bonus\ GPG : \mu_{bonus} = \frac{\mu_{bonus\_male} - \mu_{bonus\_female}}{\mu_{bonus\_male}} * 100$$

$$proportion\ of\ males\ (females)\ receiving\ bonus : Prop_{rec\_bonus} = \frac{\sum_{gender\_employees\_rec\_bonus}}{\sum_{total\_gender\_employees}} * 100$$

$$proportion\ of\ males\ (females)\ in\ each\ pay\ quartile : Prop_{by\_quart} = \frac{\sum_{gender\_employees\_in\_quartile}}{\sum_{total\_gender\_employees}} * 100$$

According to the website, 77% of organizations reporting in 2017/2018 had a gender pay gap favoring men. Since, due to COVID-19 pandemic, the GPG collection has been suspended for 2020 year, this project will concentrate on more complete and publicly available 2019 and 2018 datasets.

ADDITIONAL DATA

To determine the gender of companies' "C-level" executives, an excellent dataset from UK Office for National Statistics was utilized.

Additionally, the industry data from UK Companies House was used to convert 5-digit numeric UK Standard Industrial Classification SIC codes of economic activities to more verbose and descriptive codes. The following 21 industry sections representing more generic grouping are of specific interest:

Table 1: UK SIC Industry Sections

Code	Description
Section A	Agriculture, Forestry and Fishing
Section B	Mining and Quarrying
Section C	Manufacturing
Section D	Electricity, gas, steam and air conditioning supply
Section E	Water supply, sewerage, waste management and remediation activities
Section F	Construction
Section G	Wholesale and retail trade; repair of motor vehicles and motorcycles
Section H	Transportation and storage
Section I	Accommodation and food service activities
Section J	Information and communication
Section K	Financial and insurance activities
Section L	Real estate activities
Section M	Professional, scientific and technical activities
Section N	Administrative and support service activities
Section O	Public administration and defence; compulsory social security
Section P	Education
Section Q	Human health and social work activities
Section R	Arts, entertainment and recreation
Section S	Other service activities
Section T	Activities of households as employers; undifferentiated goods- and services-producing activities of households for own use
Section U	Activities of extraterritorial organisations and bodies

OVERVIEW AND GOALS

More insight could be gained from the available data set by analyzing the data by industry, company size and CEO gender. The gender pay gap will be projected using a gamut of models and the best performing model will be highlighted. When trying to predict a numeric value, the residuals are important sources of information. Residuals are computed as the observed value minus the predicted value (i.e., $y_i - \hat{y}_i$). When predicting numeric values, the root mean squared error (RMSE) is commonly used to evaluate models. RMSE is interpreted as how far, on average, the residuals are from zero and is calculated as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{u,i}(\hat{y}_{u,i} - y_{u,i})^2}$$

Similar to other R-squared like measures such as Mean Absolute (MAE) or Mean Bias (MBE) Errors, RMSE is a *negatively-oriented* metric, i.e., the lower values are better. However, due to squared errors, RMSE tends to give more weight to large errors relative to the others and therefore, is more sensitive to outliers. The RMSE is always larger or equal to MAE for a sample size n as in:

$$MAE \leq RMSE \leq \sqrt{n}MAE$$

Another easily interpretable metric is the coefficient of determination, R^2. This value can be interpreted as the proportion of the information in the data that is explained by the model. Practically speaking R^2 is a measure of correlation rather then accuracy.

Assuming that collected GPG data point are i.i.d. and the residuals have a theoretical mean of zero with constant variance of σ^2, the expectation of MSE could be decomposed as:

$$E[MSE] = \sigma^2 + ModelBias^2 + ModelVariance$$

The first term σ^2 represents irreducible variance that cannot be eliminated by modeling. The *squared model bias* indicates how close the functional form of the model is able to represent the true relationship between predictors and the response. The *model variance* is self explanatory. There exists a **variance bias trade-off** between the last two terms of the equation, as simpler models with low model variance tend to under-fit the true relationship while more flexible but complex models tend to over-fit due to lower model bias but high model variance.

DATA CLEANING

Extract: First Look

The first 6 rows of the initial raw data set, and its summary are displayed in Tables 2 and 3 below. We observe that:

- *EmployerName*, *companyNumber* and *CurrentName* are mutually exclusive identifier fields, so we will just keep *EmployerName* for simplicity.
- *Address* and *CompanyLinkToGPGInfo* might only be temporarily useful for grouping by *EmployerName* and should be safe to exclude from the scope of predictors going forward.
- *SicCodes* is a supplementary variable to derive industry classification by *section* and is not useful (occurring with a very low frequency) as a predictor by itself.
- *DiffMeanHourlyPercent* and *DiffMedianHourlyPercent* are two **response** variables to be analyzed and modeled. They are likely to be highly correlated so it would sufficient to pick just one as our GPG proxy. We will decide which is more suitable based on the histogram and correlation matrix.
- *DiffMeanBonusPercent*, *DiffMedianBonusPercent*, *MaleBonusPercent*, *FemaleBonusPercent* - all *bonus* related variables should be very industry specific and subject to a large reporting error. Therefore, it might be prudent to avoid them completely to reduce modeling noise.
- *MaleLowerQurtile* through *FemaleTopQuartile* should be useful predictors re-engineered into another variables as will be seen below.
- *ResponsiblePerson* is generally a top level company officer. It might be useful to derive their gender and apply as a predictor.
- *EmployerSize* is a predictor that can potentially show GPG correlation with the company size.

Table 2: Original Data: first 6 rows

EmployerName	Address	CompanyNumber	SicCodes	DiffMeanHourlyPercent	DiffMedianHourlyPercent	DiffMeanBonusPercent
'Prifysgol Aberystwyth' And 'Aberystwyth University'	Aberystwyth University, Penglais, Ceredigion, sy23 3fl	RC000641	NA	11.5	10.3	NA
10 Trinity Square Hotel Limited	5 Market Yard Mews, 194-204 Bermondsey Street, London, United Kingdom, SE1 3TQ	08064685	82990	8.7	10.3	29.6
1life Management Solutions Limited	Ldh House St Ives Business Park, Parsons Green, St. Ives, Cambridgeshire, PE27 4AA	02566586	93110, 93130, 93290	11.0	-0.5	81.5
1st Choice Staff Recruitment Limited	1ST CHOICE RECRUITMENT, 8 St. Loyes Street, Bedford, MK40 1EP	07972006	78109	-2.3	0.0	-114.8
1st Home Care Ltd.	14b Dickson Street, Elgin Industrial Estate, Dunfermline, Fife, Scotland, KY12 7SN	SC272838	86900, 88100	-2.0	0.5	NA
23.5 Degrees Limited	Unit 3 Hedge End Retail Park, Charles Watts Way, Hedge End, Southampton, Hampshire, England, SO30 4RT	08014079	56103	10.0	0.0	79.0

DiffMedianBonusPercent	MaleBonusPercent	FemaleBonusPercent	MaleLowerQuartile	FemaleLowerQuartile	MaleLowerMiddleQuartile	FemaleLowerMiddleQuartile	MaleUpperMiddleQuartile	FemaleUpperMiddleQuartile	MaleTopQuartile	FemaleTopQuartile
NA	0.0	0.0	53.0	47.0	41.0	59.0	40.0	60.0	62.0	38.0
54.5	90.5	90.5	47.9	52.1	56.3	43.7	78.9	21.1	66.7	33.3
94.2	10.0	11.4	49.0	51.0	35.3	64.7	42.3	57.7	44.2	55.8
-249.3	1.1	0.4	50.8	49.2	67.7	32.3	62.9	37.1	50.0	50.0
NA	0.0	0.0	10.0	90.0	8.0	92.0	9.0	91.0	9.0	91.0
35.0	4.0	2.0	32.0	68.0	28.0	72.0	30.0	70.0	31.0	69.0

CompanyLinkToGPGInfo	ResponsiblePerson	EmployerSize	CurrentName
https://www.aber.ac.uk/en/equality/genderpaygapreporting2019/ https://www.aber.ac.uk/cy/equality/gender-pay-gap/	Elizabeth Treasure (Vice-Chancellor)	1000 to 4999	'PRIFYSGOL ABERYSTWYTH' AND 'ABERYSTWYTH UNIVERSITY'
NA	Linda Stigter (Director of People and Culture)	250 to 499	10 TRINITY SQUARE HOTEL LIMITED
https://www.1life.co.uk/gender-pay-gap	Mark Braithwaite (Managing Director)	250 to 499	1LIFE MANAGEMENT SOLUTIONS LIMITED
https://www.1stchoice.net/gender-pay-gap-report-2019/	Gill Knight (MD)	250 to 499	1ST CHOICE STAFF RECRUITMENT LIMITED
https://reallifeoptions.org/	Ian Hardcastle (Chief Operating Officer)	250 to 499	1ST HOME CARE LTD.
https://www.23-5degrees.com/gender-pay-gap	Luca Contardo (CFO)	500 to 999	23.5 DEGREES LIMITED

Table 3: Original Data: Summary

EmployerName	Address	CompanyNumber	SicCodes	DiffMeanHourlyPercent	DiffMedianHourlyPercent	DiffMeanBonusPercent
Length:5965	Length:5965	Length:5965	Length:5965	Min. :-499.90	Min. :-134.00	Min. :-3631.00
Class :character	Class :character	Class :character	Class :character	1st Qu.: 5.50	1st Qu.: 2.00	1st Qu.: 11.60
Mode :character	Mode :character	Mode :character	Mode :character	Median : 13.70	Median : 10.60	Median : 36.00
NA	NA	NA	NA	Mean : 14.12	Mean : 12.83	Mean : 27.62
NA	NA	NA	NA	3rd Qu.: 22.30	3rd Qu.: 22.00	3rd Qu.: 58.00
NA	NA	NA	NA	Max. : 100.00	Max. : 100.00	Max. : 100.00
NA	NA	NA	NA	NA	NA	NA's :1527

DiffMedianBonusPercent	MaleBonusPercent	FemaleBonusPercent	MaleLowerQuartile	FemaleLowerQuartile	MaleLowerMiddleQuartile	FemaleLowerMiddleQuartile	MaleUpperMiddleQuartile
Min. :-2300.00	Min. : 0.00	Min. : 0.00	Min. : 0.00	Min. : 0.00	Min. : 0.00	Min. : 0.00	Min. : 0.0
1st Qu.: 0.00	1st Qu.: 0.00	1st Qu.: 0.00	1st Qu.: 26.00	1st Qu.: 38.10	1st Qu.: 28.40	1st Qu.: 30.80	1st Qu.: 33.0
Median : 20.50	Median : 19.70	Median : 17.90	Median : 42.00	Median : 58.00	Median : 46.20	Median : 53.80	Median : 52.0
Mean : 10.95	Mean : 37.13	Mean : 35.85	Mean : 44.46	Mean : 55.54	Mean : 48.87	Mean : 51.13	Mean : 53.5
3rd Qu.: 45.33	3rd Qu.: 78.60	3rd Qu.: 77.40	3rd Qu.: 61.90	3rd Qu.: 74.00	3rd Qu.: 69.20	3rd Qu.: 71.60	3rd Qu.: 76.0
Max. : 100.00	Max. :100.00	Max. :100.00	Max. :100.00	Max. :100.00	Max. :100.00	Max. :100.00	Max. :100.0
NA's :1529	NA	NA	NA	NA	NA	NA	NA

FemaleUpperMiddleQuartile	MaleTopQuartile	FemaleTopQuartile	CompanyLinkToGPGInfo	ResponsiblePerson	EmployerSize	CurrentName
Min. : 0.0	Min. : 0.00	Min. : 0.00	Length:5965	Length:5965	Length:5965	Length:5965
1st Qu.: 24.0	1st Qu.: 40.20	1st Qu.: 20.00	Class :character	Class :character	Class :character	Class :character
Median : 48.0	Median : 60.80	Median : 39.20	Mode :character	Mode :character	Mode :character	Mode :character
Mean : 46.5	Mean : 59.29	Mean : 40.71	NA	NA	NA	NA
3rd Qu.: 67.0	3rd Qu.: 80.00	3rd Qu.: 59.80	NA	NA	NA	NA
Max. :100.0	Max. :100.00	Max. :100.00	NA	NA	NA	NA

Transformation: Employee Size to Numeric

To begin it is useful to convert the character column with ***EmployerSize*** ranges to a new column ***Min Employees*** as seen in Table 4 below:

Table 4: Transformation: Employee Size to Numeric

EmployerSize	MinEmployees
Less than 250	1
250 to 499	250
500 to 999	500
1000 to 4999	1000
5000 to 19,999	5000
20,000 or more	20000

Transformation: Reduce with Grouping by 'ResponsiblePerson'

The raw dataset has multiple records for the same entity for instance satellite or regional offices reporting separately as per Table 5.

Table 5: Same Company - Multiple Records

EmployerName	Address	CompanyNumber	DiffMeanHourlyPercent	DiffMedianHourlyPercent	ResponsiblePerson	EmployerSize	CurrentName
Capita Business Services Ltd	65 Gresham Street, London, England, EC2V 7NQ	02299747	34.2	28.8	Will Serle (Chief People Officer)	20,000 or more	CAPITA BUSINESS SERVICES LTD
Capita Customer Management Limited	65 Gresham Street, London, England, EC2V 7NQ	01336850	9.7	2.2	Will Serle (Chief People Officer)	5000 to 19,999	CAPITA CUSTOMER MANAGEMENT LIMITED
Capita Employee Benefits (Consulting) Limited	65 Gresham Street, London, England, EC2V 7NQ	01860772	37.4	44.2	Will Serle (Chief People Officer)	250 to 499	CAPITA EMPLOYEE BENEFITS (CONSULTING) LIMITED
Capita Employee Benefits Limited	65 Gresham Street, London, England, EC2V 7NQ	02260524	20.9	13.6	Will Serle (Chief People Officer)	1000 to 4999	CAPITA EMPLOYEE BENEFITS LIMITED
Capita It Services Limited	Pavilion Building Ellismuir Way, Tannochside Park, Uddingston, Glasgow, G71 5PW	SC045439	25.2	32.8	Will Serle (Chief People Officer)	500 to 999	CAPITA IT SERVICES LIMITED
Capita Life & Pensions Regulated Services Limited	65 Gresham Street, London, England, EC2V 7NQ	02424853	16.9	15.6	Will Serle (Chief People Officer)	1000 to 4999	CAPITA LIFE & PENSIONS REGULATED SERVICES LIMITED
Capita Life & Pensions Services Limited	65 Gresham Street, London, England, EC2V 7NQ	04359665	35.5	36.2	Will Serle (Chief People Officer)	250 to 499	CAPITA LIFE & PENSIONS SERVICES LIMITED
Capita Managed It Solutions Limited	Hillview House, 61 Church Road, Newtownabbey, Co Antrim, BT36 7LQ	NI032979	20.5	17.7	Will Serle (Chief People Officer)	500 to 999	CAPITA MANAGED IT SOLUTIONS LIMITED
Capita Plc	65 Gresham Street, London, England, EC2V 7NQ	02081330	30.9	31.7	Will Serle (Chief People Officer)	250 to 499	CAPITA PLC
Capita Property And Infrastructure Limited	65 Gresham Street, London, England, EC2V 7NQ	02018542	23.5	21.6	Will Serle (Chief People Officer)	1000 to 4999	CAPITA PROPERTY AND INFRASTRUCTURE LIMITED
Capita Resourcing Limited	65 Gresham Street, London, England, EC2V 7NQ	03949686	25.9	40.8	Will Serle (Chief People Officer)	5000 to 19,999	CAPITA RESOURCING LIMITED
Capita Retail Financial Services Limited	65 Gresham Street, London, England, EC2V 7NQ	05296886	1.5	6.2	Will Serle (Chief People Officer)	1000 to 4999	CAPITA RETAIL FINANCIAL SERVICES LIMITED
Capita Secure Information Solutions Limited	65 Gresham Street, London, England, EC2V 7NQ	01593831	15.2	13.7	Will Serle (Chief People Officer)	1000 to 4999	CAPITA SECURE INFORMATION SOLUTIONS LIMITED
Capita Travel And Events Limited	65 Gresham Street, London, England, EC2V 7NQ	01094729	36.2	30.4	Will Serle (Chief People Officer)	500 to 999	CAPITA TRAVEL AND EVENTS LIMITED
Entrust Support Services Limited	The Riverway Centre, Riverway, Stafford, United Kingdom, ST16 3TH	04440463	-3.1	7.0	Will Serle (Chief People Officer)	500 to 999	ENTRUST SUPPORT SERVICES LIMITED
Fera Science Limited	65 Gresham Street, London, England, EC2V 7NQ	09413107	13.6	6.6	Will Serle (Chief People Officer)	250 to 499	FERA SCIENCE LIMITED
G L Hearn Limited	65 Gresham Street, London, England, EC2V 7NQ	03798877	43.6	41.1	Will Serle (Chief People Officer)	250 to 499	G L HEARN LIMITED
Optima Legal Services Limited	Hepworth House, Clarpit Lane, Leeds, LS2 8AE	05781608	17.3	5.0	Will Serle (Chief People Officer)	250 to 499	OPTIMA LEGAL SERVICES LIMITED
Re (Regional Enterprise) Limited	65 Gresham Street, London, England, EC2V 7NQ	08615172	17.2	11.1	Will Serle (Chief People Officer)	250 to 499	RE (REGIONAL ENTERPRISE) LIMITED
Tascor Services Limited	65 Gresham Street, London, England, EC2V 7NQ	02057887	12.4	1.0	Will Serle (Chief People Officer)	250 to 499	TASCOR SERVICES LIMITED
Trustmarque Solutions Limited	65 Gresham Street, London, England, EC2V 7NQ	02183240	21.2	25.2	Will Serle (Chief People Officer)	250 to 499	TRUSTMARQUE SOLUTIONS LIMITED
Updata Infrastructure (Uk) Limited	65 Gresham Street, London, England, EC2V 7NQ	06957593	15.7	19.9	Will Serle (Chief People Officer)	250 to 499	UPDATA INFRASTRUCTURE (UK) LIMITED
Voice Marketing Limited	65 Gresham Street, London, England, EC2V 7NQ	05820091	-0.5	0.2	Will Serle (Chief People Officer)	500 to 999	VOICE MARKETING LIMITED
Western Mortgage Services Limited	65 Gresham Street, London, England, EC2V 7NQ	03191608	20.8	9.8	Will Serle (Chief People Officer)	500 to 999	WESTERN MORTGAGE SERVICES LIMITED

Therefore, we can reduce the original set with 5,965 rows and 22 columns grouping by ***ResponsiblePerson***, taking means of GPG fields, and taking a row with max Employee Count in case they are different.

The resulting dataset has 3,800 rows and 20 columns with numeric ***MinEmployees*** field from prior transformation. Adding back 1,300 rows of the original set with NA for ***ResponsiblePerson*** , we are still able to accomplish 15% reduction in size without any loss of information.

Transformation: Extract 'Job Function' to get 'C-level' executives

Each ***ResponsiblePerson*** such as *Aamir Khalid (Chief Executive)*, can be converted into two separate fields: ***FirstName*** and ***JobFunction*** using this Regex. Of these two supplementary fields ***FirstName*** will imply a person's gender while ***JobFunction*** should define which ***ResponsiblePerson*** could be considered a power executive that hypothetically could be called a **C-level** executive.

To separate 'C-level' from middle management another Regex comes in handy. It should pull ***JobFunction*** with words in *(CEO, CFO, Chief, Head, Chair, Chairman, Founder, Partner, President)* while excluding the words such as *(Hr, Staff, Human, People, Personnel, Talent, Sales, Compensation, Benefits, Reward, Compliance, Vice)*. This way, *Vice President* is not being considered a 'C-level' type, but *President* would.

Table 6: Transformation: Extracting JobFunction

title	CLevel
Chief Executive	TRUE
Regional Vice President London & General Manager The Carlton Tower Jumeirah	FALSE
Chief Executive Officer	TRUE
Managing Director	FALSE
Finance Director	FALSE
Reward Manager	FALSE
Human Resources Director	FALSE
International Finance Operations Manager	FALSE
Chief Financial Controller	TRUE
Head	TRUE

Transformation: Gender from 'Name'

To imply the gender from the first name, the initial idea to harvest the gender from a baby name website such as nameberry.com did not prove fruitful. It worked exceptionally well for gender unique names, but for the names applicable to both genders such as Robin the website did not have counts of the name being likelier female or male. Instead, the dataset from UK Office for National Statistics was exploited to estimate gender proportions based on counts for UK since 1996.

Table 7: Transformation: Extracting FirstName

name	Gender	prop
Aamir	male	1.0000000
Aaron	male	1.0000000
Abdallah	male	1.0000000
Abigail	female	1.0000000
Adam	male	1.0000000
Adelle	female	1.0000000
Adil	male	1.0000000
Adrian	male	0.9993287
Ahmad	male	1.0000000
Ahmed	male	1.0000000

Transformation: Verbose Industry sections

The industry sections were scraped from SIC codes website to make use of various SIC industry codes reported by company:

Table 8: Transformation: Extracting Industry Sections

Code	Description	section
Section A	Agriculture, Forestry and Fishing	Agriculture, Forestry and Fishing
01110	Growing of cereals (except rice), leguminous crops and oil seeds	Agriculture, Forestry and Fishing
01120	Growing of rice	Agriculture, Forestry and Fishing
01130	Growing of vegetables and melons, roots and tubers	Agriculture, Forestry and Fishing
01140	Growing of sugar cane	Agriculture, Forestry and Fishing
01150	Growing of tobacco	Agriculture, Forestry and Fishing
01160	Growing of fibre crops	Agriculture, Forestry and Fishing
01190	Growing of other non-perennial crops	Agriculture, Forestry and Fishing
01210	Growing of grapes	Agriculture, Forestry and Fishing
01220	Growing of tropical and subtropical fruits	Agriculture, Forestry and Fishing

Load

A sample few rows of of the **cleaned** dataset shown in the Table 9 below before FEATURE ENGINEERING transformations has taken place. As discussed at the beginning of this section, we are to remove all supplementary or redundant fields such as *SicCodes*, all bonus variables, *Weblink*, *ResponsiblePerson*, *name*, *tittle*, *CLevel*, *Description*, before adding new variables.

Table 9: Transformed Data: first 6 rows

EmployerName	SicCodes	DiffMeanHourlyPercent	DiffMedianHourlyPercent	DiffMeanBonusPercent	DiffMedianBonusPercent
'Prifysgol Aberystwyth' And 'Aberystwyth University'	NA	11.50	10.30	NA	NA
10 Trinity Square Hotel Limited	82990	8.70	10.30	29.6	54.5
1life Management Solutions Limited	93110, 93130, 93290	11.00	-0.50	81.5	94.2
1st Choice Staff Recruitment Limited	78109	-2.30	0.00	-114.8	-249.3
1st Home Care Ltd., Real Life Options	86900, 88100, 87300, 88100	1.05	0.75	NA	NA
23.5 Degrees Limited	56103	10.00	0.00	79.0	35.0

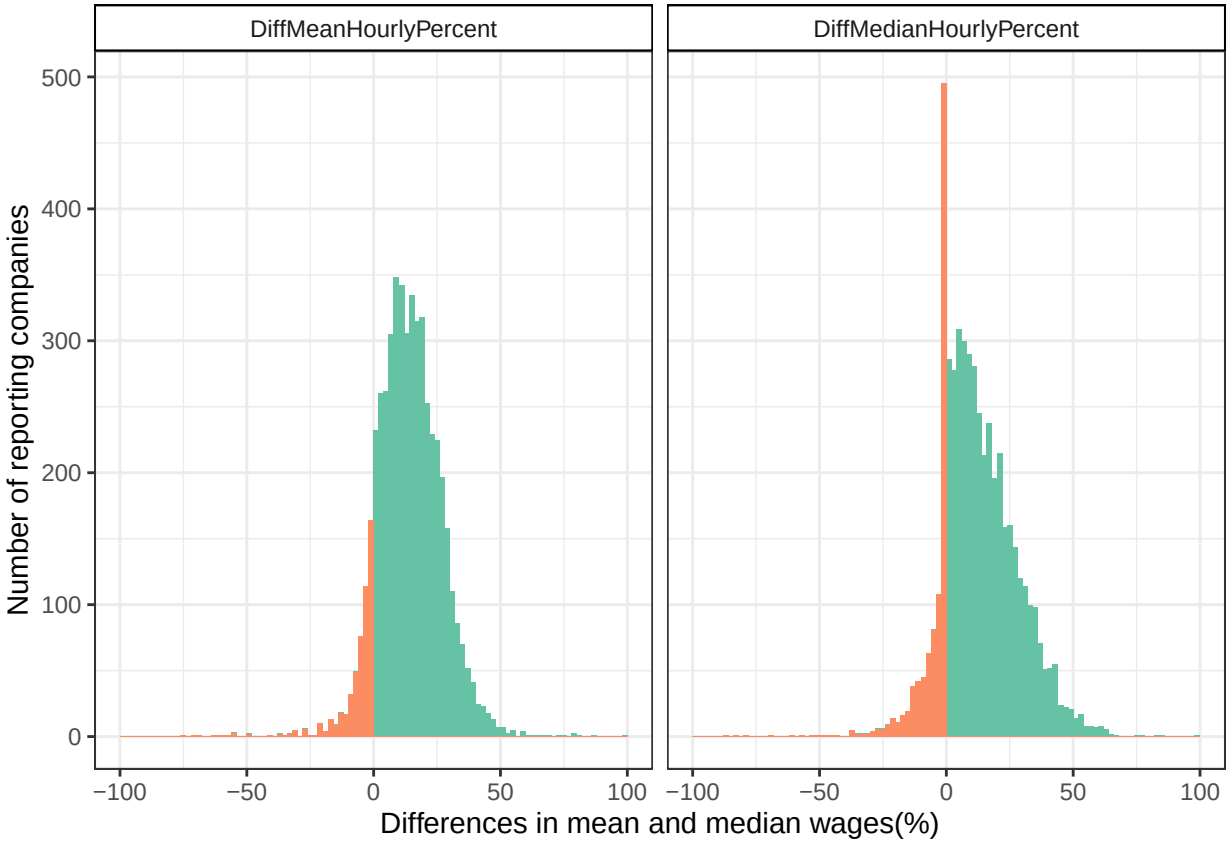
MaleBonusPercent	FemaleBonusPercent	MaleLowerQuartile	FemaleLowerQuartile	MaleLowerMiddleQuartile	FemaleLowerMiddleQuartile	MaleUpperMiddleQuartile
0.0	0.0	53.0	47.0	41.0	59.0	40.0
90.5	90.5	47.9	52.1	56.3	43.7	78.9
10.0	11.4	49.0	51.0	35.3	64.7	42.3
1.1	0.4	50.8	49.2	67.7	32.3	62.9
0.0	0.0	15.0	85.0	14.5	85.5	14.5
4.0	2.0	32.0	68.0	28.0	72.0	30.0

FemaleUpperMiddleQuartile	MaleTopQuartile	FemaleTopQuartile	WebLink	ResponsiblePerson	MinEmployees
60.0	62.0	38.0	https://www.aber.ac.uk/en/equality/genderpaygapreporting2019/ https://www.aber.ac.uk/cy/equality/gender-pay-gap/	Elizabeth Treasure (Vice-Chancellor)	1000
21.1	66.7	33.3	NA	Linda Stigter (Director of People and Culture)	250
57.7	44.2	55.8	https://www.1life.co.uk/gender-pay-gap	Mark Braithwaite (Managing Director)	250
37.1	50.0	50.0	https://www.1stchoice.net/gender-pay-gap-report-2019/	Gill Knight (MD)	250
85.5	15.0	85.0	https://reallifeoptions.org/	Ian Hardcastle (Chief Operating Officer)	1000
70.0	31.0	69.0	https://www.23-5degrees.com/gender-pay-gap	Luca Contardo (CFO)	500

name	title	CLevel	Gender	Description	section
Elizabeth	Vice Chancellor	FALSE	female	NA	NA
Linda	Director Of People And Culture	FALSE	female	Other business support service activities n.e.c.	Administrative and support service activities
Mark	Managing Director	FALSE	male	Operation of sports facilities, Fitness facilities, Other amusement and recreation activities n.e.c.	Arts, entertainment and recreation
Gill	Md	FALSE	NA	Other activities of employment placement agencies	Administrative and support service activities
Ian	Chief Operating Officer	TRUE	male	Other human health activities, NA, Social work activities without accommodation for the elderly and disabled	Human health and social work activities
Luca	Cfo	TRUE	male	Take-away food shops and mobile food stands	Accommodation and food service activities

DATA EXPLORATION & VISUALIZATION

As both ***median*** and ***mean*** values are reported, the plot below summarizes all the data with positive values for the companies where GPG is in men favor (81% of companies for the median and 89% for the mean), while negative values imply GPG in women favor. It can be observed that unlike the mean, which transitions smoothly, the median changes abruptly and is a bit sticky as a measure. The large number of median zeros can be explained by the companies that have a large proportion of their workforce (over 50 percent) paid the exact same rate, e.g. in retail. It is then not difficult to also have both the median male and the median female employee fall into that category, which would result in the median gap being zero, but not the mean. Given that both mean and median are highly correlated (see the correlation matrix further down), it might be prudent to use mean as the sole response variable to predict.

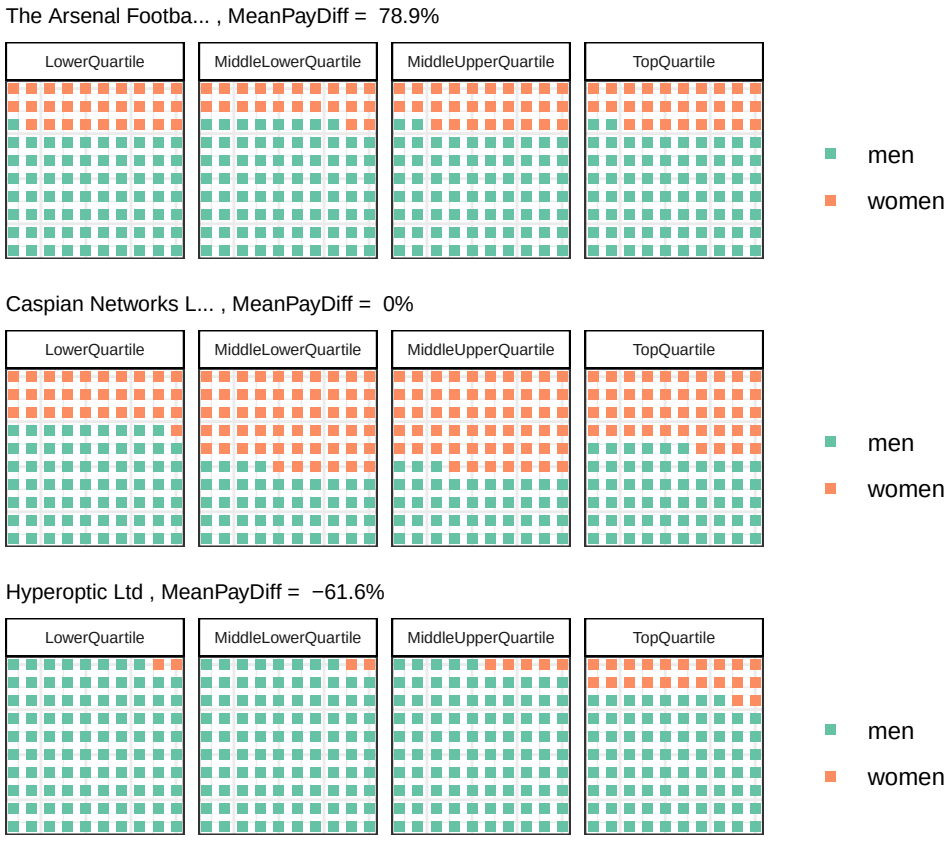


Moving on, we can inspect another piece of available information, namely the reporting of gender proportion in each pay quartile. Let's take a slice of companies with at least a ***thousand*** employees on the payroll and pick three of them: with maximum **positive** mean hourly pay difference (representing largest pay gap favoring men), a zero mean ("equality"), and maximum **negative** (largest pay gap favoring women). *Note that each quartile is plotted with 100 dots for visual percent representation, and does not represent an actual employee count.*

1. **Arsenal Football Club** shows quite a large gap. However, allowing for a fact that football in UK has been dominated by men where former players are likely to occupy more senior and managerial roles, it is not all that surprising. What is mildly surprising is to see a percentage of women represented in Top Pay Quartile.

2. **Caspian** shows approximately equal representation of both female and male employees in each quartile. The tilt is in favor of men in Top Quartile is likely offset by more women in the two Middle Quartiles.

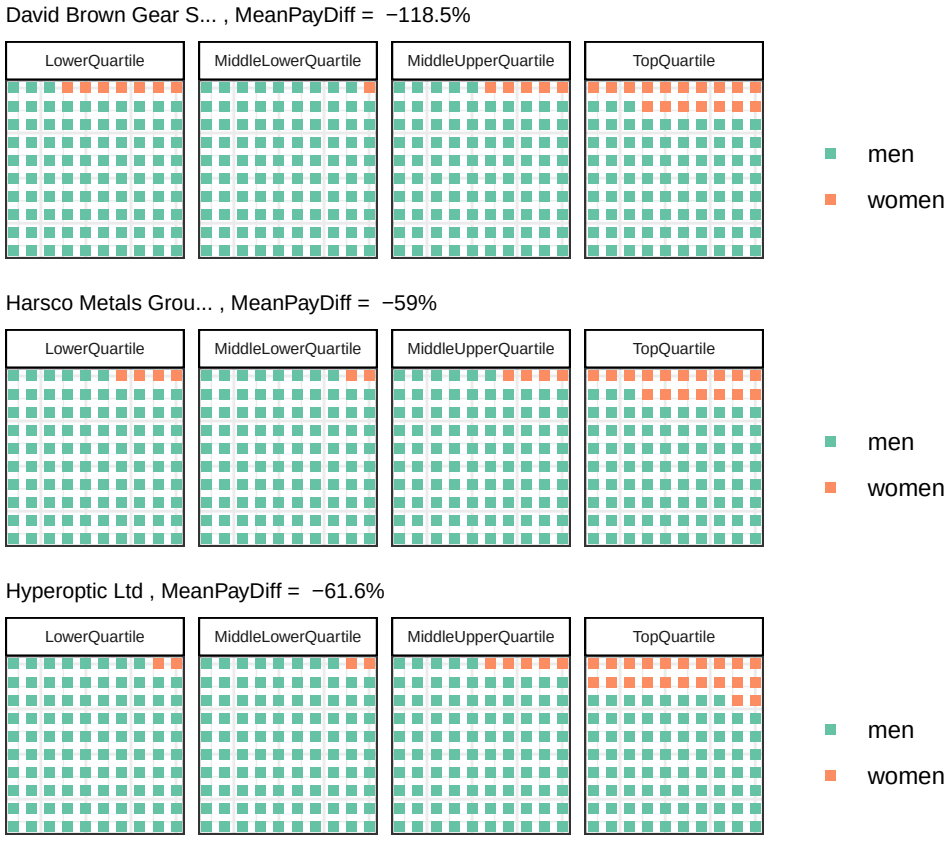
3. **Hyperoptic Ltd** is the most interesting. It has a very few female employees all together with the largest portion of them are in the top quartile (still only ~20%). What is really driving the -61.6% GPG in women favor then? Let's look at the executive team to gain more insight. One third (CEO and two more) out of nine total executives are female, similar to what Top Quartile shows. It is not impossible that female CEO takes home the lion share of compensation comparing to other executives. If that were the case, it would not represent a win for women. On the contrary, one may argue that even Arsenal looks better employing more women across the board.



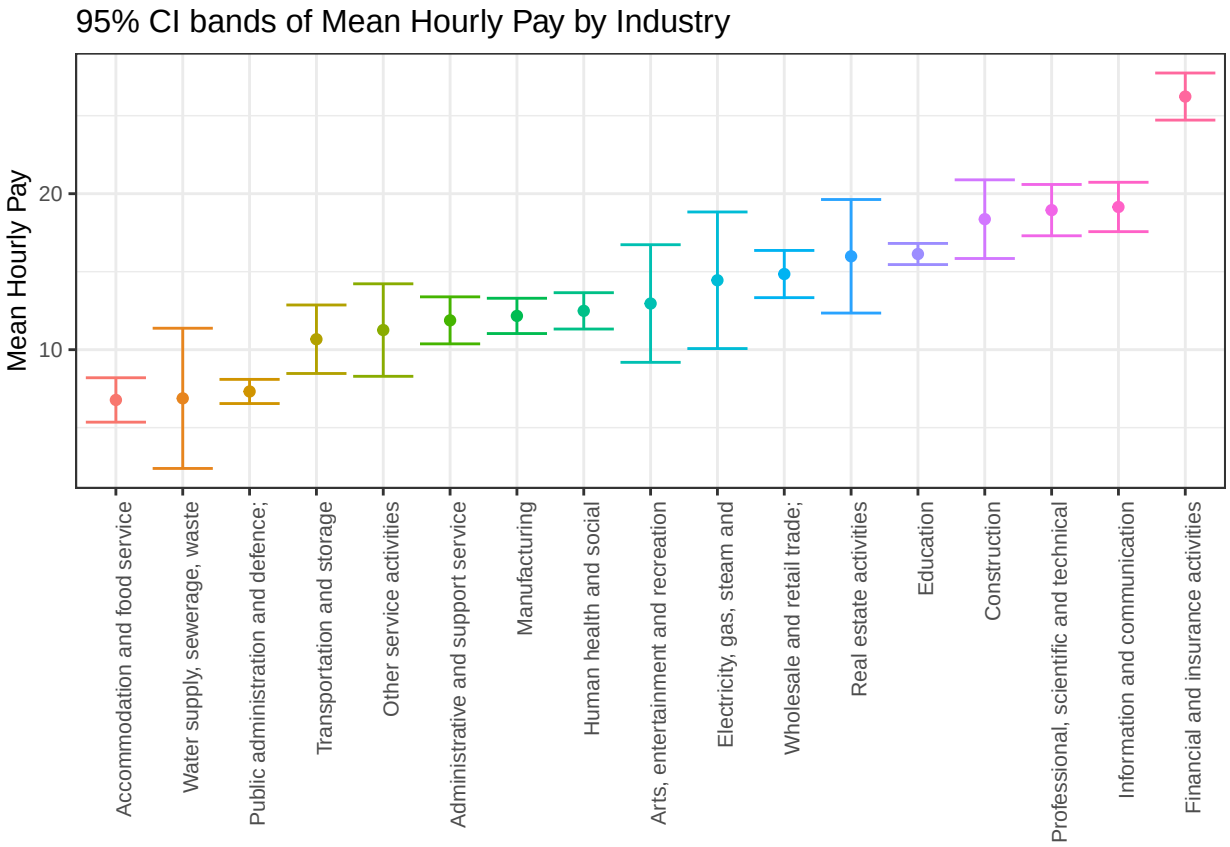
To see what is really happening with **Hyperoptic Ltd**, we sampled other two firms with 250 - 500 and 500 - 1000 employees respectively where Mean Pay Difference is the largest negative number in the group (GPG favoring women). We see that both **David Brown** and **Harsco Metals** pay quartile distributions of women to men look strikingly similar to **Hyperoptic Ltd.** The prevalence of male employees could be explained by industry as both are metals and heavy machinery. While the executive team for David Brown is not available, it is hard to make sense of the negative Mean Pay by looking at **Harsco Metals** executive team with only female as HR head to 9 men in other leadership roles. As counter-intuitive as that may seem at first, let's think it through: high participation for men implies their representation on the whole range of pay rates, from the lowest to the highest incomes. Lack of flexibility in the market makes it less enticing for low-skilled women to join or not worthwhile for employers to take on the costs of hiring women, or both. This means the women who are working at this companies are more likely to be *better skilled* and therefore better paid on average, which shifts their wage distribution higher than if their participation had been at the same level as that of men's.

Scaling this concept up to country levels, research has found a strong correlation between GPG and labor participation rates, i.e., countries with low levels of female labor participation are also the ones where GPG is the lowest. Here is the quote from one such study: *"The gender pay gap in some countries is distorted, even in favour of women in some countries (Bahrain, Costa Rica) due to the small proportion of women in the working population's formal economy. For example, in the Middle East (where only about one third of*

women participate in the workforce), the gender pay gap is minus 40 per cent in Bahrain due to the greater representation of women in ***elite roles*** *in the labour market compared with the small number of women represented in the total workforce, while* ***the average earnings figure for men is pulled down*** *by the greater representation of men in the overall labour market, and as a consequence also in the lower-paid roles."*



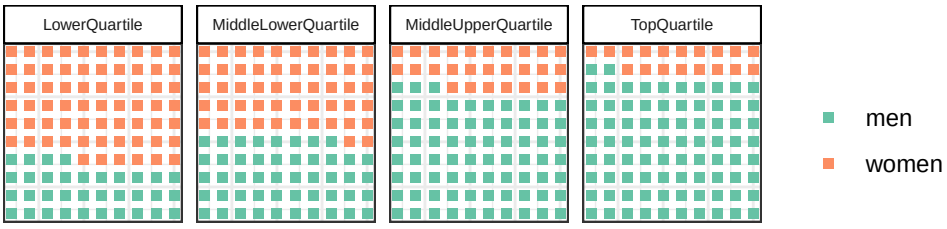
To continue data exploration, it is informative to consider the breakdown by industry in the next plot. The records with missing industry sector are excluded as well as industries with less than 10 companies. The bands of Mean Hourly Pay Difference are sorted by average of Hourly Pay Difference (dots on the plot) and bars indicating 2 standard error deviation from the mean. It can be observed that GPG is relatively low in accommodation, food , utilities, and public sectors and gradually increasing in "higher" paid industries with **Finance and Insurance** having the highest GPG.



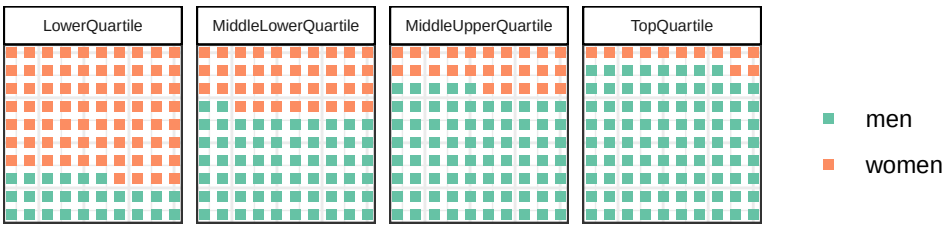
Naturally, it would be now quite interesting to inspect the pay quartiles of 3 sampled companies with at least 250 employees that belong to **Finance and Insurance** or the industry with the highest GPG. All three have a very similar quartile distribution pattern as well Mean Pay Differences of around 50%. There are more women than men employed in the lowest paid quartile while it gradually reverses for other three quartiles with the least proportion of women in the top pay quartile. Compared to the companies with negative Mean Pay Difference (GPG in favor of women) where there were a larger proportion of women in top paying quartiles, this is the opposite. Again, we can hypothesize that high GPG in **Finance and Insurance** is due to having more men in senior positions. This is most likely due to a number of historical, educational and industry specific factors and not a result of "unequal pay for the same work".

Labor participation rates could be a critical factor impacting GPG. Let's take our thinking a step further and consider the following potential measure to reduce if not reverse GPG for the sampled finance and insurance companies below. Presumably, firing all women in the lower pay quartile and replacing them with men would turn the tables and should have an effect similar to **Hyperoptic Ltd** and drastically improve GPG metric by itself. However, let us ask ourselves if that really would be a win for women…

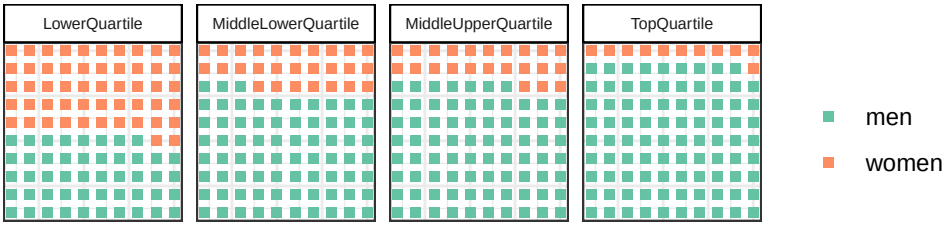
Brown Shipley & Co... , MeanPayDiff = 50.5%



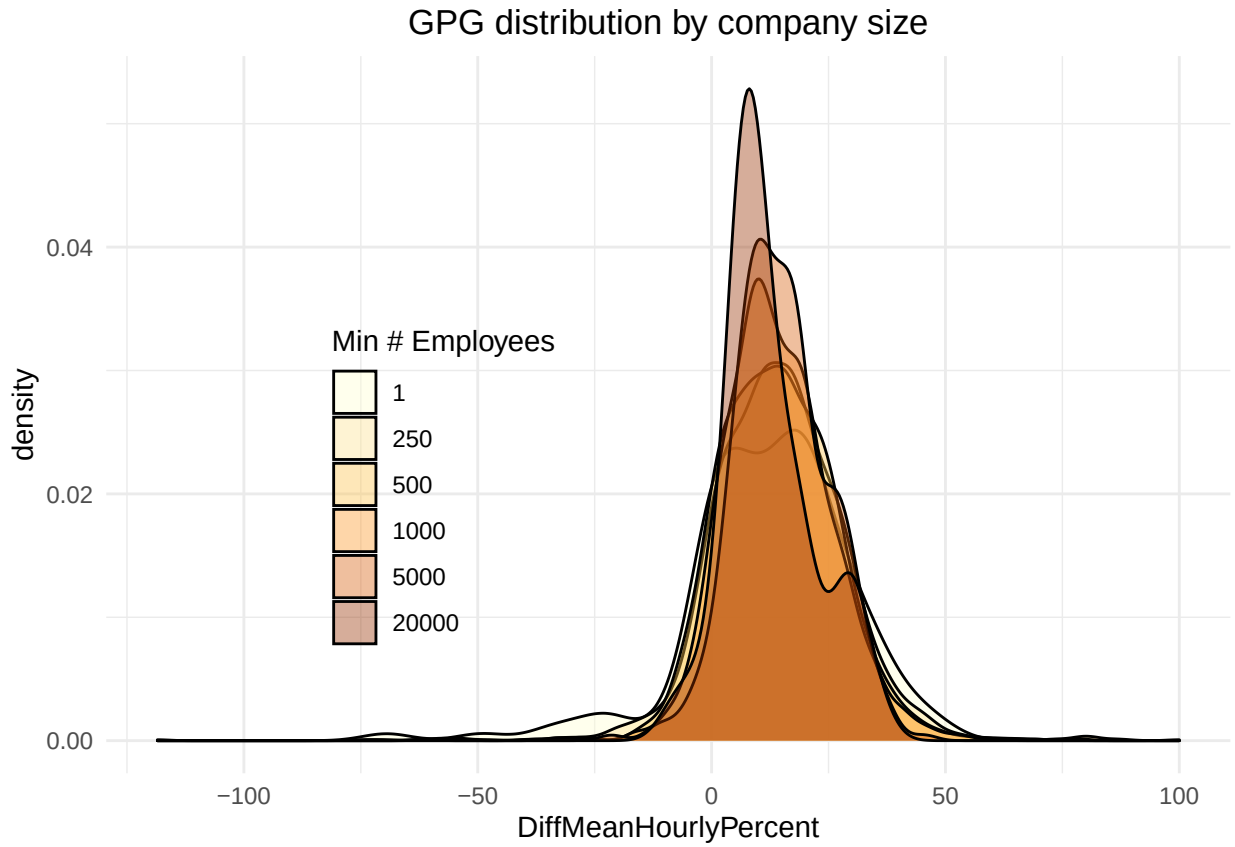
N. M. Rothschild &... , MeanPayDiff = 51.3%



Nomura Internation... , MeanPayDiff = 47.3%

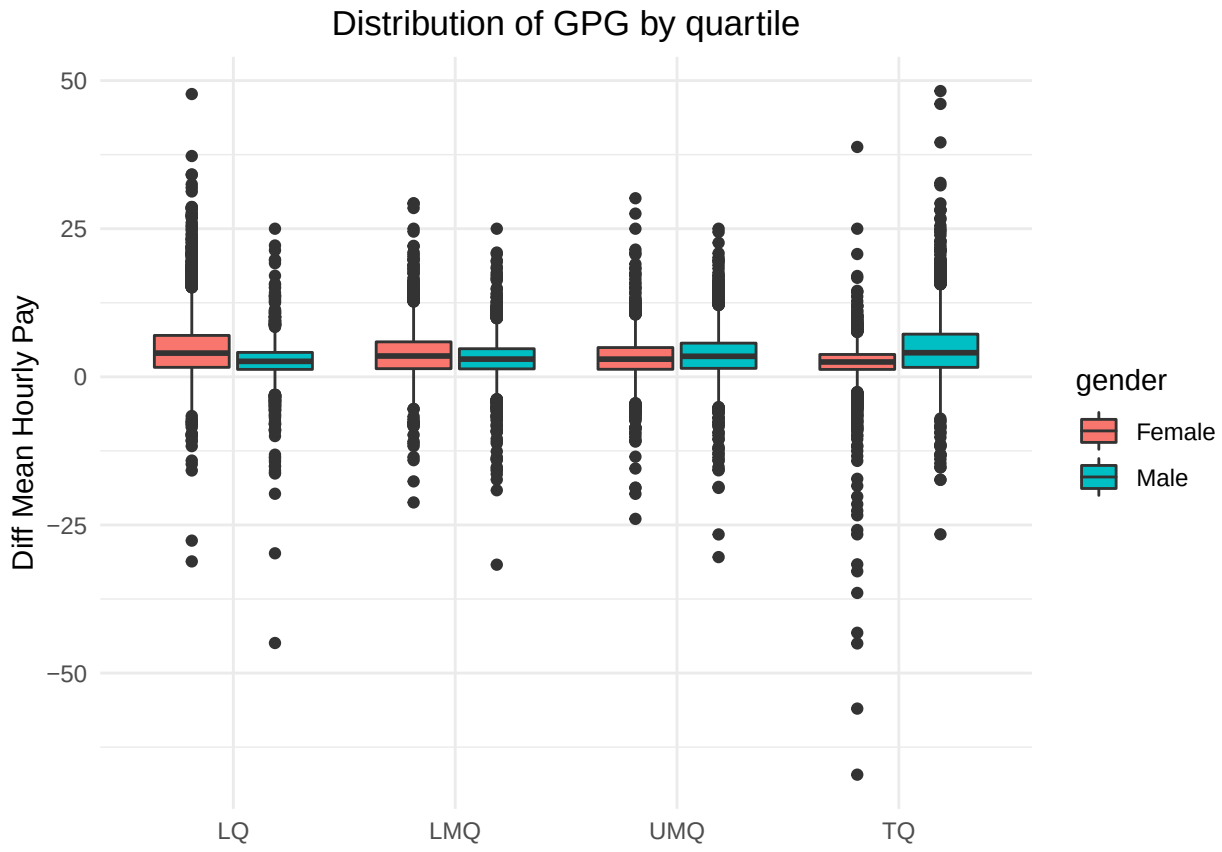


Exploring the company size potential predictive power, we observe from the below plot that it does not exhibit any meaningful relationships. The response is right skewed and distributed fairly uniformly in relation to the company size:



On the other hand, our derived predictor for CEO gender appears to show some correlation:

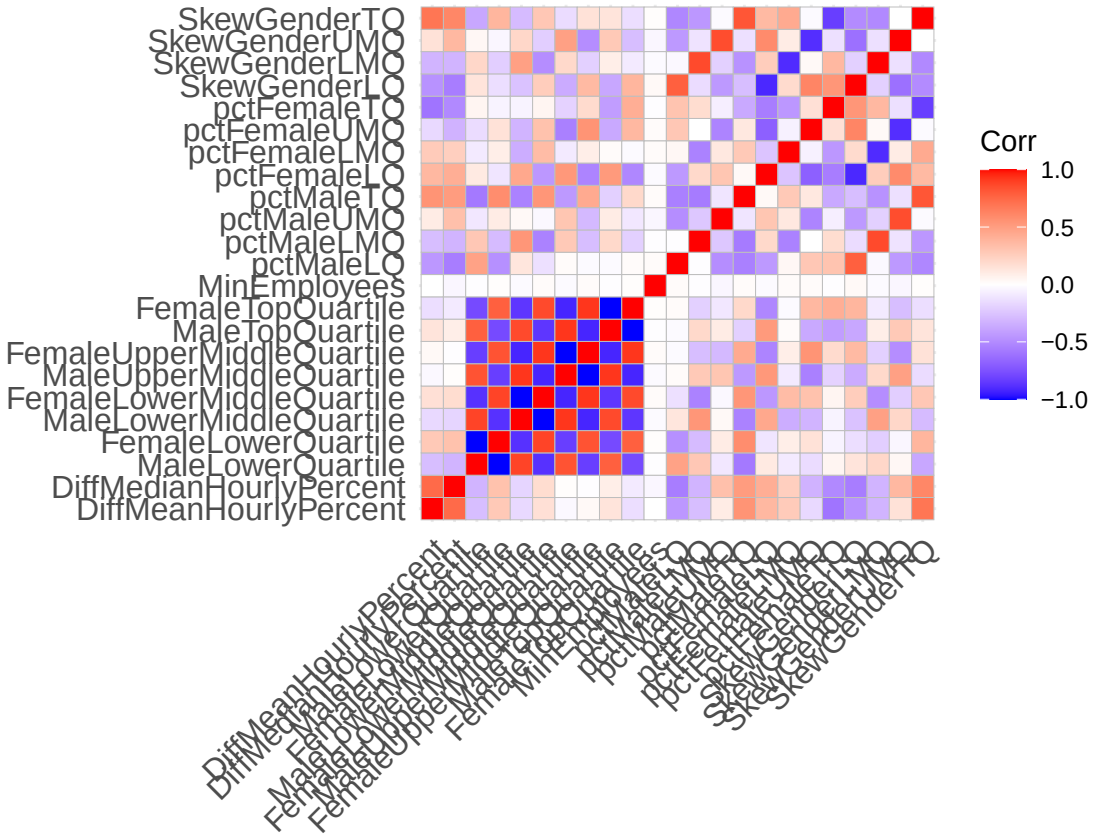
A larger volume and higher median in two lower quartiles for female and the same for males in the top two quartiles could be observed on the below box-plot of pay quartile ratios. Noticeably, the dispersion is more pronounced **in the bottom and top quartiles** than in the middle two. In addition, the presence of numerous *outliers* in the whiskers indicates the noisiness of the data which will complicate our predictive ability.



Finally, the correlation matrix for all numerical variables should prove instructive of variables to remove. In general, there are good reasons to avoid data with highly correlated predictors. First, redundant predictors frequently add more complexity to the model than information they provide to the model. It could also be computationally costly to have more variables. Finally, using highly correlated predictors in techniques like linear regression can result in highly unstable models, numerical errors, and degraded predictive performance. All these reasons are critical to the Neural Networks models which will be reviewed in more details in the MODELING section.

It can be observed that

- Two response variables, mean and median are highly correlated so leaving mean and removing median is confirmed to make sense.
- Original female and male percent quartiles have weak correlation with the responses. However, our derived skew predictors show stronger correlation. Moreover, the correlation goes from high negative to high positive by skew quartile.
- the company size shows a weak relation to the response which is most likely due to scaling differences (company size in 1000 while the gap in %)



FEATURE ENGINEERING

"If you torture the data long enough, it will confess."
–Ronald H. Coase

Adding Variables: Quartile Skew

The way the predictors are encoded can have a significant impact on model performance. For example, using combinations of predictors can sometimes be more effective than using the individual values: the ratio of two predictors may be more effective than using two independent predictors. Often the most effective encoding of the data is suggested by the our understanding of the problem and thus is not derived from any mathematical technique. Here, we engineer a new set of variables from percent quartiles by estimating the ratio of n^{th} quartile over total percent by gender and calculating the skew of male vs female by quartile:

$$pct\ Male\ n^{th}\ Quartile\ = \frac{male\ n^{th}Q}{maleLowerQ + maleLowerMiddleQ + maleUpperMiddleQ + maleTopQ}$$

$$Skew\ Gender\ n^{th}\ Quartile\ =\ pctMale\ n^{th}Q - pctFemale\ n^{th}Q$$

Adding Variables: Converting factors to numeric

When a predictor is categorical such as gender or ethnicity, it is beneficial to decompose the predictor into a set of binary variables. For example, for our gender predictor with $value = female$, we would generate two "dummy" variables $gender.male = 0$ and $gender.female = 1$. Another categorical variable for the *industry section* would generate 20 variables for each section.

Removing variables

Of the dummy variables generated for the *industry section*, there are a few with a frequency that is severely disproportionate. These are called **near-zero** variance predictors and should be removed.

Transformation: Center & Scale

To center a predictor variable, the average predictor value is subtracted from all the values. As a result of centering, the predictor has a zero mean. Similarly, to scale the data, each value of the predictor variable is divided by its standard deviation. Scaling the data coerce the values to have a common standard deviation of one. These manipulations are used to improve the numerical stability of data at the expense of interpretability of the individual values as the data is no longer in the original units.

Transformation: Variable Reduction

There are potential advantages to removing predictors prior to modeling:

- Fewer predictors means decreased computational time and complexity.
- If two predictors are highly correlated, this implies that they are measuring the same underlying information. Removing one should not compromise the performance of the model and might lead to a more parsimonious and interpretable model.
- Some models can be crippled by predictors with degenerate distributions. In such cases, model performance and/or stability can be improved significantly without the problematic variables.

The variable reduction is accomplished via PCA from *caret PreProcess* function. While the function also has the ability to impute the missing data using various methods such a *knn*, the sample size is large enough so that we could just drop the missing data without much loss of information.

Transformation: Remove Skewness

This is to transform the predictors so that the probability of falling on either side of the distribution's mean is roughly equal. R *caret* package uses a statistical technique called "Box Cox" that determines an appropriate λ parameter (e.g. - 1 for inverse or .5 for square root transformation) using MLE. This would be applied independently for each predictor. Skewness by predictor is shown below and it can be observe that Minimum Number of Employees is very skewed.

	Skew Values
MaleLowerQuartile	0.32
FemaleLowerQuartile	-0.32
MaleLowerMiddleQuartile	0.24
FemaleLowerMiddleQuartile	-0.24
MaleUpperMiddleQuartile	0.08
FemaleUpperMiddleQuartile	-0.08
MaleTopQuartile	-0.16
FemaleTopQuartile	0.16
MinEmployees	6.69
pctMaleLQ	0.34
pctMaleLMQ	-0.12
pctMaleUMQ	0.82
pctMaleTQ	1.39
pctFemaleLQ	1.46
pctFemaleLMQ	-0.24
pctFemaleUMQ	-0.69
pctFemaleTQ	1.73
SkewGenderLQ	-0.20
SkewGenderLMQ	0.35
SkewGenderUMQ	0.19
SkewGenderTQ	-0.69

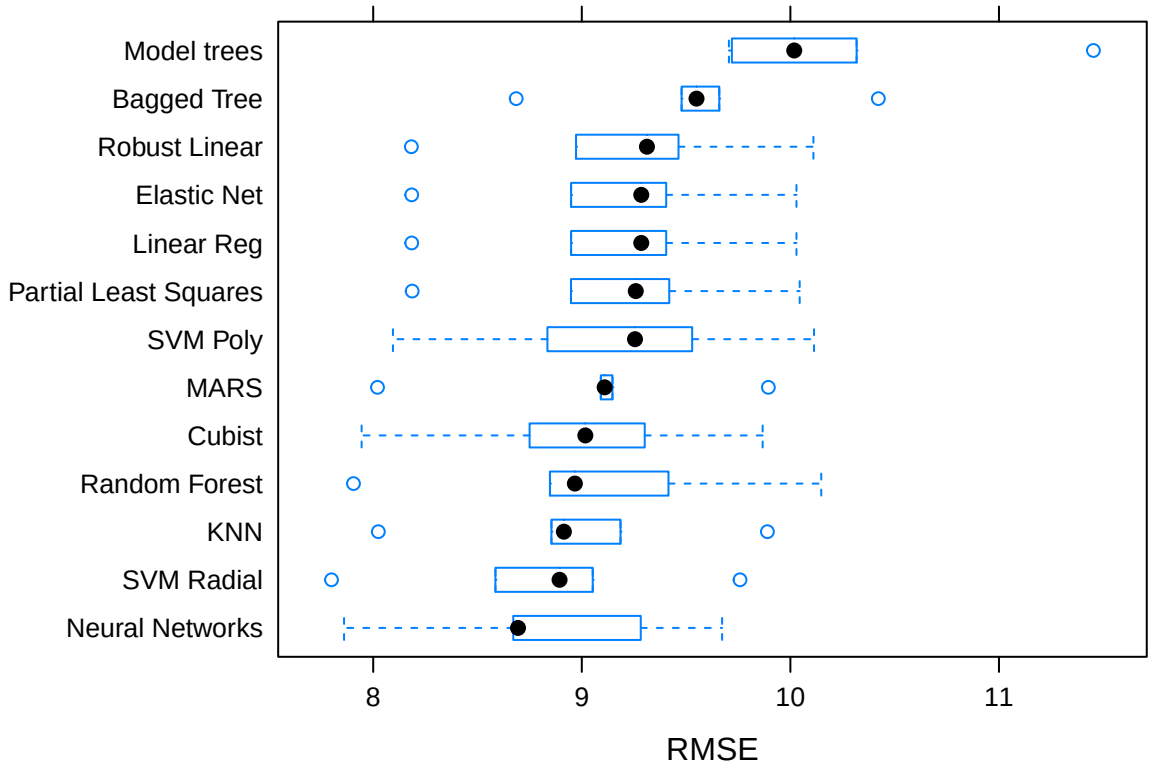
MODELLING

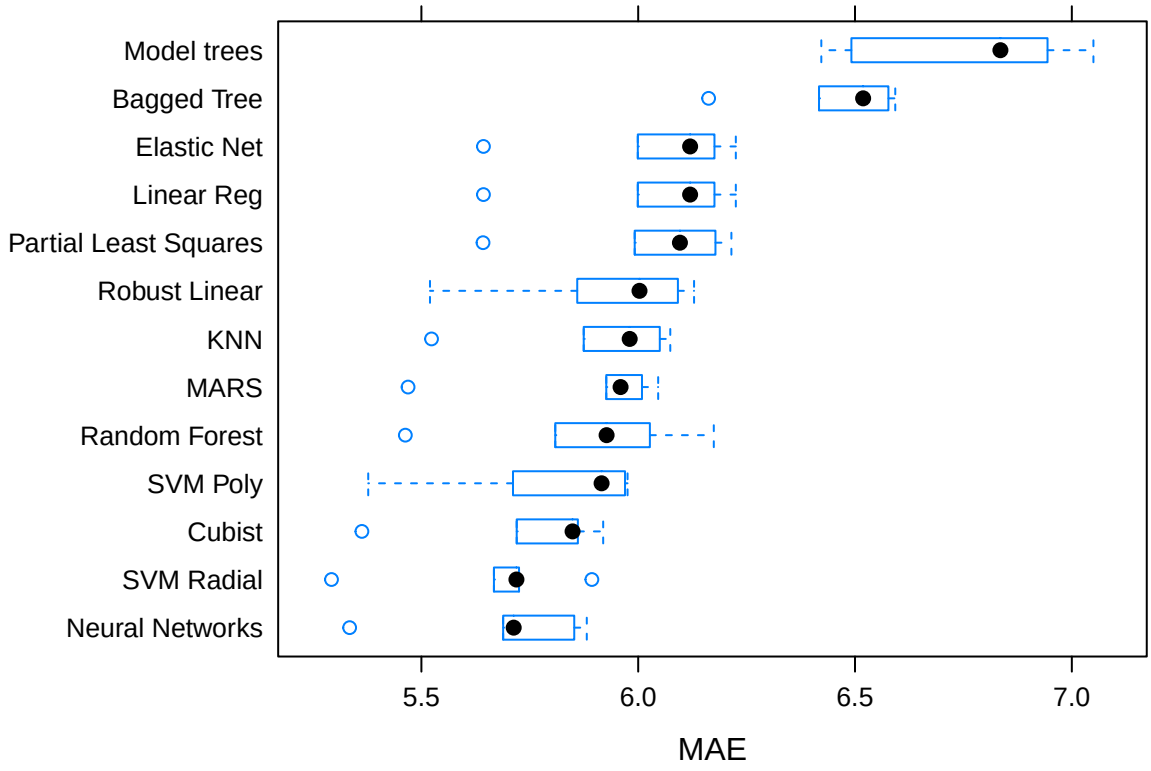
"All models are wrong, but some are useful."
–George E.P. Box

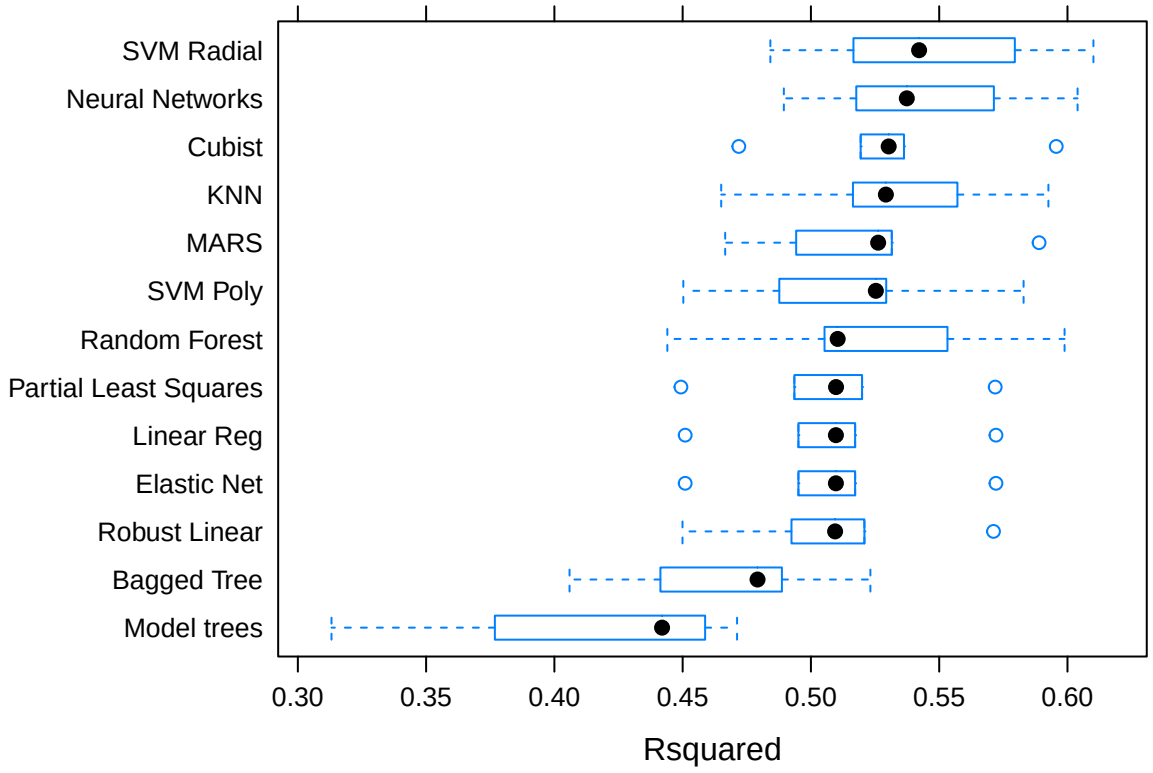
Next, we are going to select a regression (not a classification) model concerned with predicting the GPG which is a continuous variable. Most of the background on GPG has been already provided in the earlier sections.

Almost all predictive modeling techniques have tuning parameters that enable the model to flex to find the structure in the data. Hence, we must use the existing data to identify settings for the model's parameters that yield the best and most realistic predictive performance (known as model tuning). Traditionally, this has been achieved by splitting the existing data into training and test sets. We will be using 90/10 partition and 5-fold cross-validation to tune a series of models starting with more "rigid" linear models, followed by non-linear models, decision trees and rule-based models.

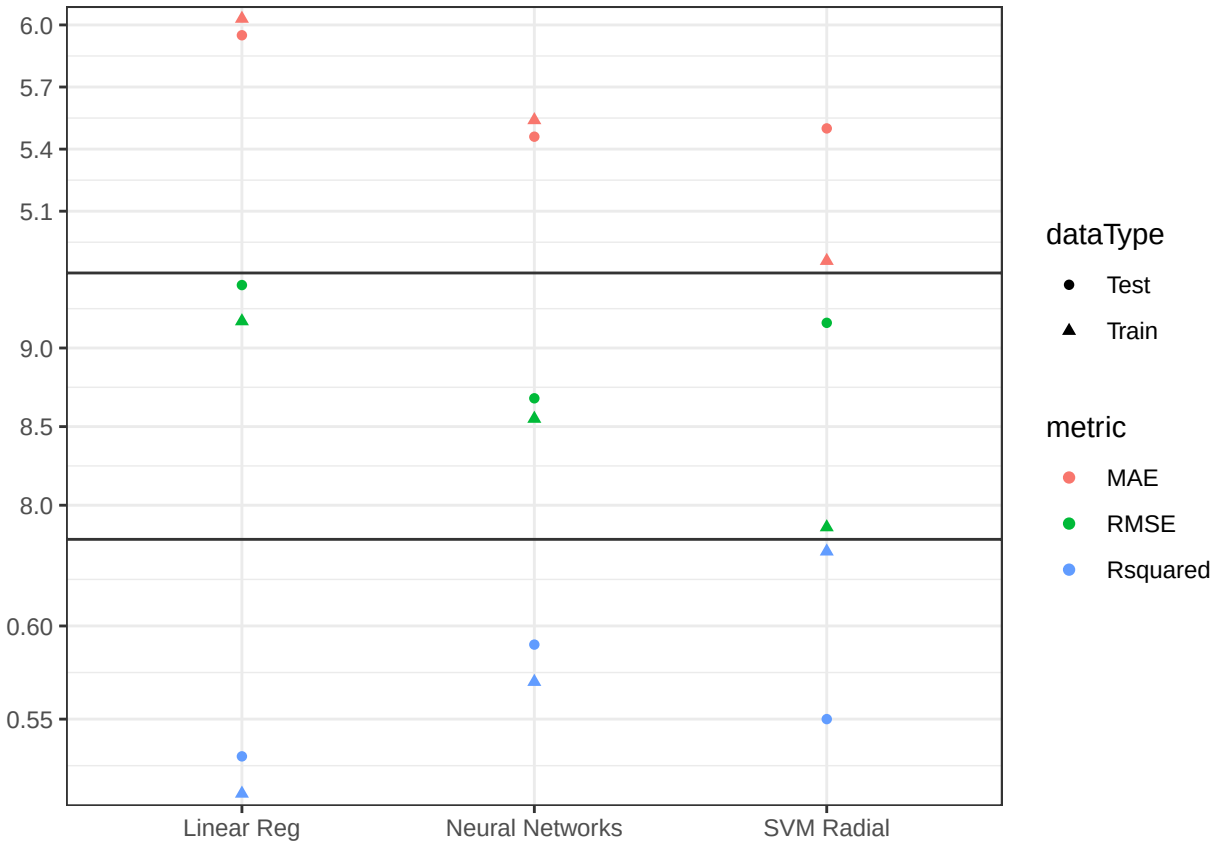
Looking at the following three consecutive box plots for each metric, the majority of trained models demonstrate surprisingly uniform results, with Support Vector Machine with Radial Kernel ("SVM") and Model Averaged Neural Network ("AvgNNet") having the slight edge over the others, while Model Trees and Bagged Tree models under-performing. It is also quite remarkable that the variations of linear models such as Partial Least Squares or Elastic Net did not do better than Ordinary Least Squares Linear Regression ("LM"). As the dataset is prone to outliers as has been shown on "Distribution of GPG by quartile" plot in DATA EXPLORATION section, MAE results are also shown as MAE is expected to be less sensitive to the outliers compared to $RMSE$. However, the MAE metrics by model are fairly consistent with $RMSE$.




Model trees
Bagged Tree
Elastic Net
Linear Reg
Partial Least Squares
Robust Linear
KNN
MARS
Random Forest
SVM Poly
Cubist
SVM Radial
Neural Networks
5.5
6.0
6.5
7.0
MAE



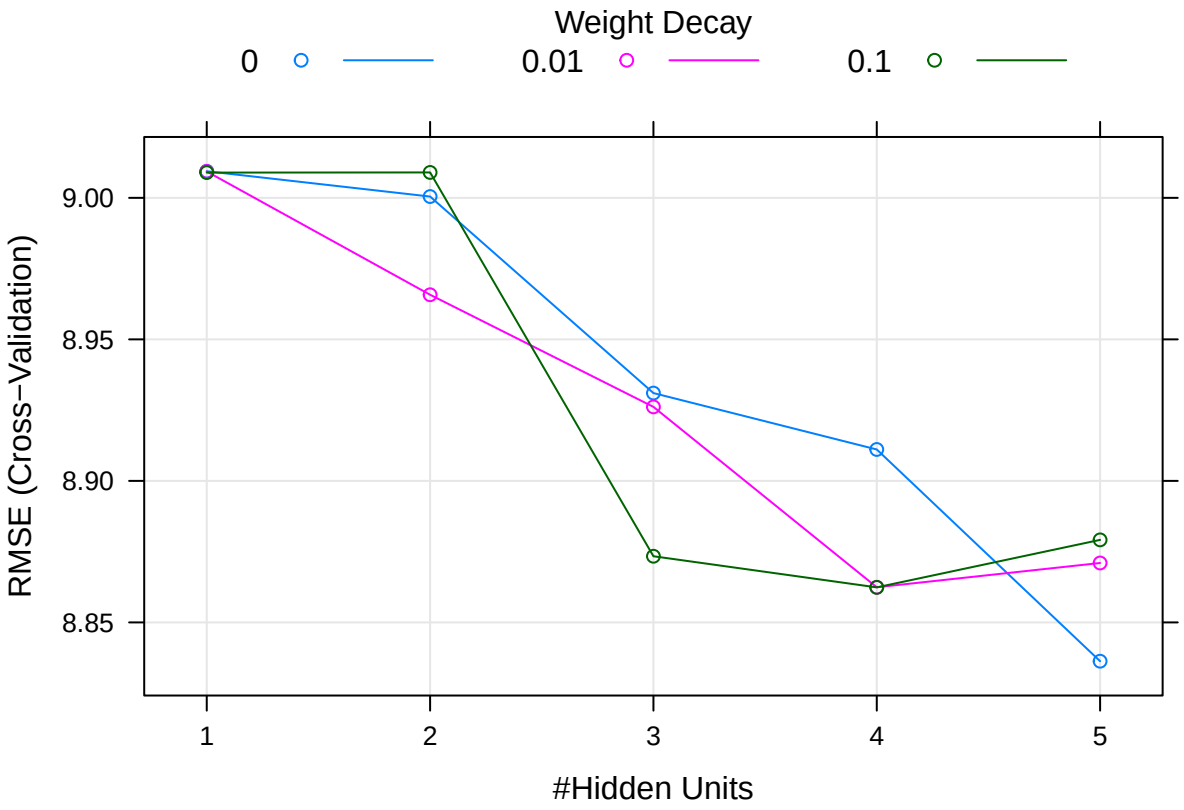
Based on the train set results, three models, **Linear Regression, Neural Network and Support Vector Machine** are preliminary selected for further assessment using the test dataset (10% of the original data). The below plot and table compare the metrics of these models predictions fitted to the test set:



	Train			Test		
model	RMSE	Rsquared	MAE	RMSE.tst	Rsquared.tst	MAE.tst
Linear Reg	**9.17**	0.51	6.03	**9.40**	0.53	5.95
Neural Networks	**8.55**	0.57	5.54	**8.68**	0.59	5.46
SVM Radial	**7.86**	0.64	4.86	**9.16**	0.55	5.50

As could also be seen from the plot, SVM model is the one poorly performed on the test data. It indicates significant over-fitting having produced a test $RMSE$ of 9.16 compared to train $RMSE$ of 7.86, or 1.3 $RMSE$ increase. It also shows a **lower** test R^2 (0.55 vs 0.64). Contrary to SVM, the Neural Networks model looks very stable on the test data, with only 0.11 $RMSE$ increase from 8.55 trained $RMSE$. It also appears to be marginally better than Linear Regression with $RMSE$ lower by about 0.7 less $RMSE$ and 6-7% larger R^2. Therefore, we declare *Model Averaged Neural Networks* a winner! Let us describe it in more details.

Neural Networks create at least one layer of go-in-between variables that cannot be observed (Hidden Units). The relationship between the original predictors and this layer of variables is non-linear. However, the response is modeled from this intermediary subset using linear relationship. Every coefficient in resulting linear combinations, β_{ij}, carries the effect of i^{th} original predictor on the j^{th} intermediary variable. For the model used here, the number of total parameters in the final prediction equation is defined as $K(P+1)+K+1$ where K is the number of intermediary variables and P is the number of predictors (5 post PCA transformation in our model). Neural Networks could grow increasingly complicated having multiple intermediary subsets of Hidden Units modeling each other. This often leads to optimization issues, model instability, and tendency to over-fit. A technique called *model averaging* uses random starting guesses to create several models and then average their results. Model averaging is known to produce a more stable prediction.

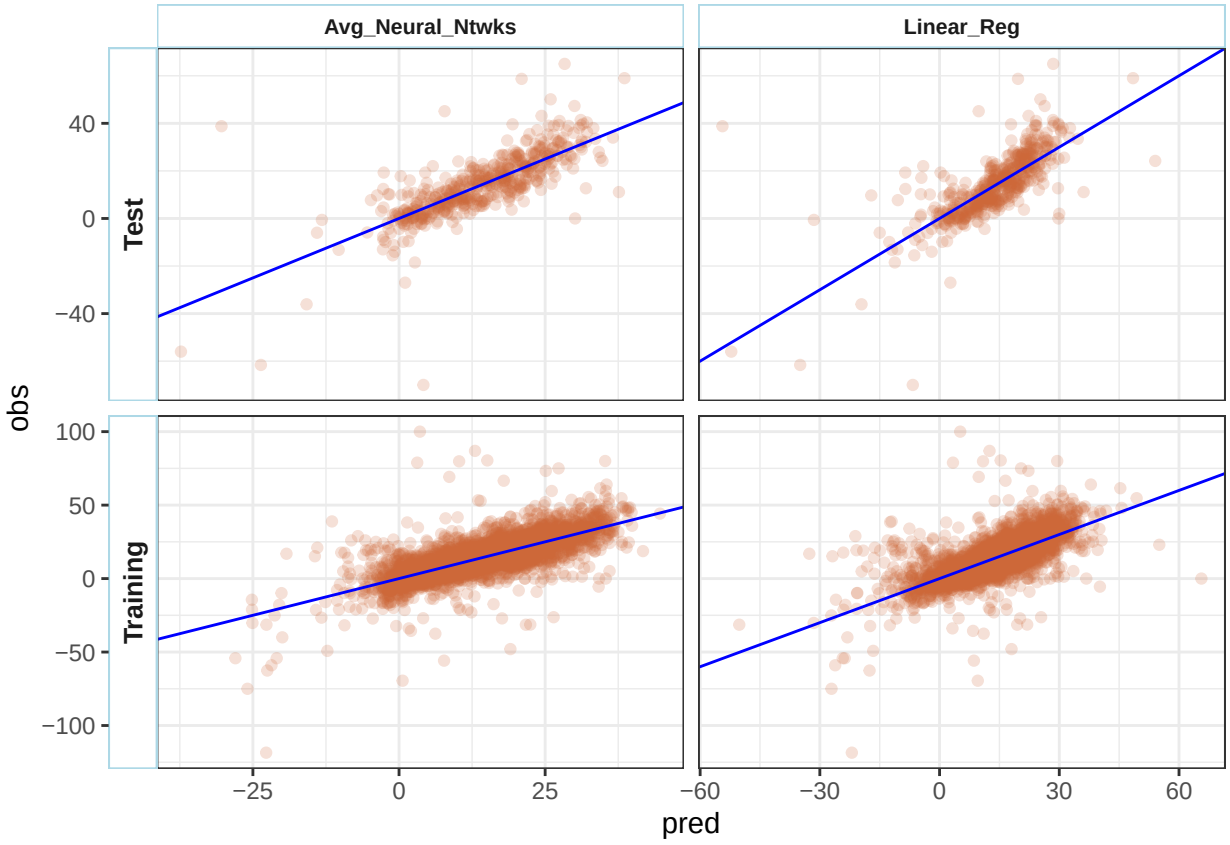


For this project, Model Averaged Neural Networks uses a single intermediary layer with the following tuning parameters:

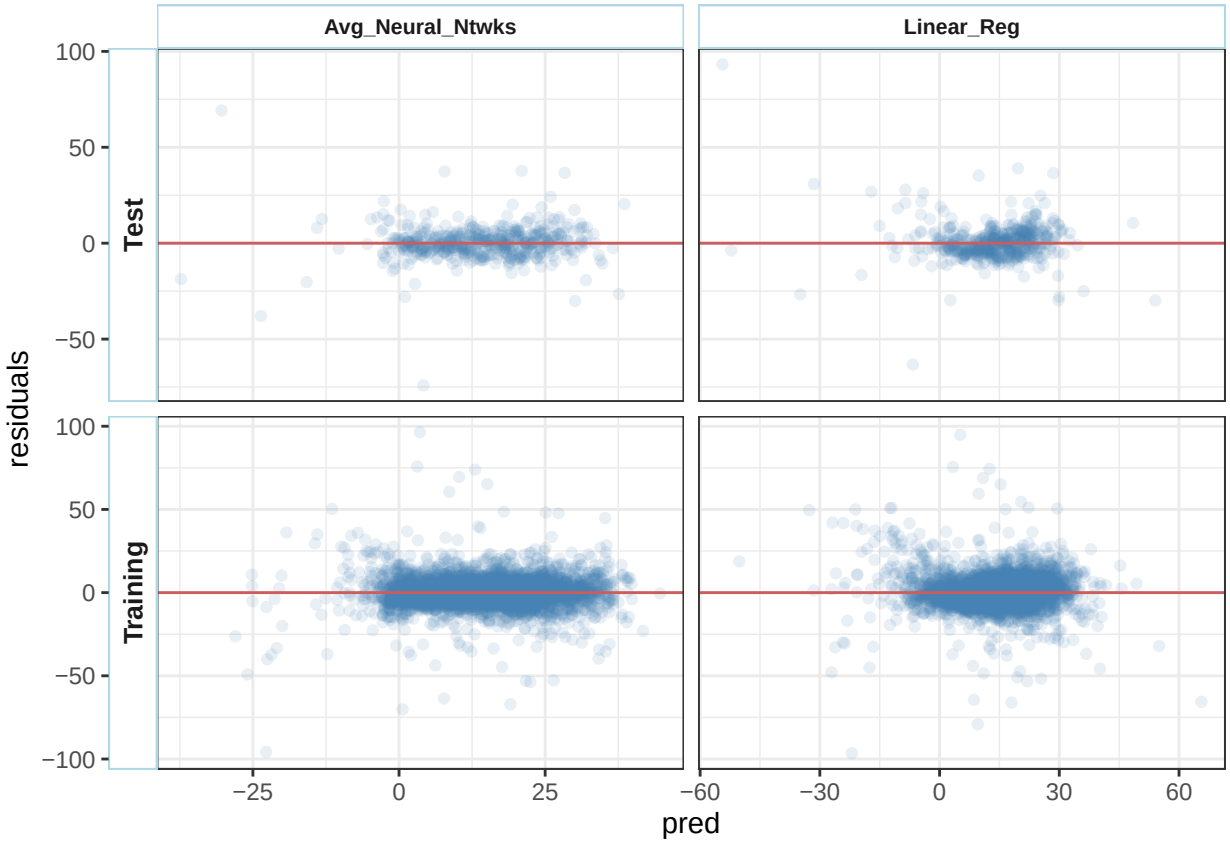
- *caret* method: 'avNNet'
- weight decay: $\lambda = (0, 0.01, 0.1)$ to penalize large regression coefficients to reduce the over-fitting
- size of neurons in the layer: 1 to 5
- maximum iteration: 500
- bag: FALSE , i.e., no bootstrapping

The optimal model per RMSE criteria plot above uses $\lambda = 0$ and 5 neurons. The fairly simplistic Neural Network model utilized only 5 PCA predictors and was not computationally expensive. However, should we not have used PCA to reduce the number of predictors or added more predictors (see CONCLUSION) it could require substantially more computational muscle.

To evaluate the quality of the models we proceed to visualize the results. The plot of **the observed against the predicted** values helps to understand how well models fit. The diagonal blue line indicates where observed and predicted values would be equal. For both remaining models, Linear Regression and Neural Network, there is no visual evidence of over- or under-fitting, i.e., there is no systematic bias in the predictions.



In addition, **the residual against the predicted** plot shows the residuals are more or less randomly scattered about red horizontal line (at zero) and there are no noticeable clusters or patterns. There is no clear winner in the two models based on these plots which look very similar.



CONCLUSION

This has been quite the journey! A number of models performed similarly to predict the GPG and **Model Averaged Neural Networks** did slightly better than others. Was it significantly better? It is not easy to decide, but given that the model is essentially a black box with a highly complex prediction algorithm, the linear regression looks just as attractive, is simpler and a lot more interpretable. Consequently, we feel that **Linear Regression** should be the model of choice for our GPG modeling.

For GPG response variable with *mean* of 14 and *standard deviation* of 13.6, the Neural Networks model achieved a test $RMSE$ of 8.68 and R^2 of 0.59. Are these good or bad metrics? As $RMSE$ measures the standard deviation of the error distribution, being smaller than a standard deviation of response is encouraging. However, the real answer is that it cannot be determined solely based on just statistical expertise and without additional competence in the subject matter. Still, based on the outside GPG research reviewed, our dataset predictor space leaves out some of most critical predictors such as *age, race, tenure, highest education level or commute time* among others. As such, it is advisable to collect more predictors to be able to explain more variation not captured in the data.

Among other future considerations for predictors, a data transformation called *spatial sign* could be performed to minimize the sensitivity to outliers. Also, given a substantial size of the data set, all observations with missing data we removed. Inputting them using *knn* or other method could potentially improve the fit.

BIBLIOGRAPHY

Irizarry, Rafael A. (2019) Introduction to Data Science : Data Analysis and Prediction Algorithms with R

Xie, Yihui; Allaire, J.J.; Grolemund, Garrett (2019) Rmarkdown: The Definitive Guide

Kuhn, Max; Johnson, Kjell (2013) "Applied Predictive Modeling", Springer http://appliedpredictivemodeling.com/

James, Gareth; Witten, Daniela; Hastie, Trevor; Tibshirani, Robert (2013) "An Introduction to Statistical Learning" , Springer

Wickham, Hadley (2009), "Elegant Graphics for Data Analysis" , Springer