Is the experience of junior doctors in England associated with hospital mortality, hospital performance or patient satisfaction?

Table of contents

1 Introduction

Facing escalating workplace pressures exacerbated by the COVID-19 pandemic and recent waves of industrial action, "junior doctors," comprising half of the UK medical workforce, confront increased dissatisfaction and burnout. A recent meta-analysis in the BMJ has demonstrated how junior doctor burnout causes career disengagement and a poorer quality of care [@hodkinson2022associations]. The authors explored the complex factors that affect doctor burnout and disengagement, which can be seen in Figure 1.

Despite these challenges, there is limited data on the direct link between doctor dissatisfaction and adverse patient outcomes at a hospital level. Using the 2022 General Medical Council

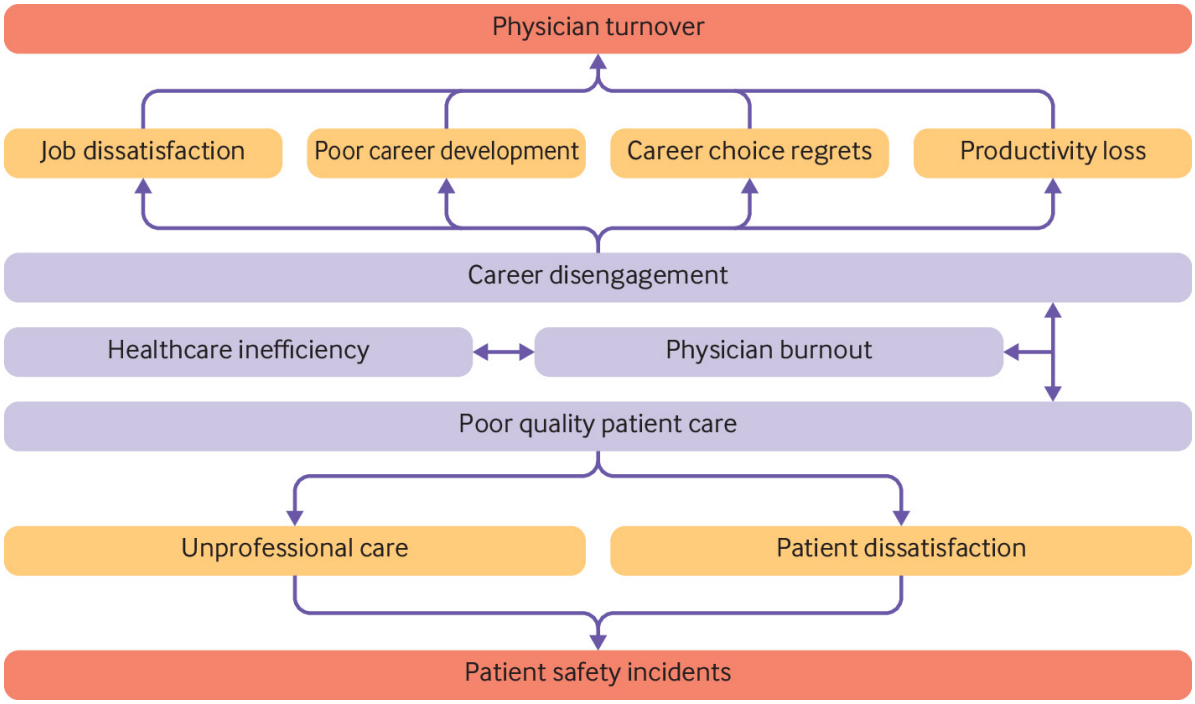


Figure 1: Figure 1

(GMC) national training survey, which assesses 18 different aspects, I conducted a comparative analysis on three metrics of performance of foundation hospital trusts in England . I examined how junior doctors, defined as all doctors below consultant level who are in a training programme, rated their working environment in terms of 'clinical supervision,' 'clinical supervision outside regular hours,' 'teamwork,' and 'rota design.' I then compared these ratings to three key hospital performance metrics for the year 2022: the Summary Hospital-level Mortality Indicator (SHMI), the latest Care Quality Commission (CQC) ratings, and the 2022 adult inpatient survey results.

Given the large number of factors that influence hospital performance, I consider the following confounders when constructing my regression models: percentage of emergency department (ED) attendances seen in less than 4 hours; percentage of occupied hospital beds; overall yearly financial performance as measured by income and expenses; number of junior doctors in a training programme; sickness absence rate; percentage of trainees who responded to the national training survey; and the stability index for turnover all staff except junior doctors, defined by $\frac{(Joiners-Leavers)}{(Joiners+Leavers)}$.

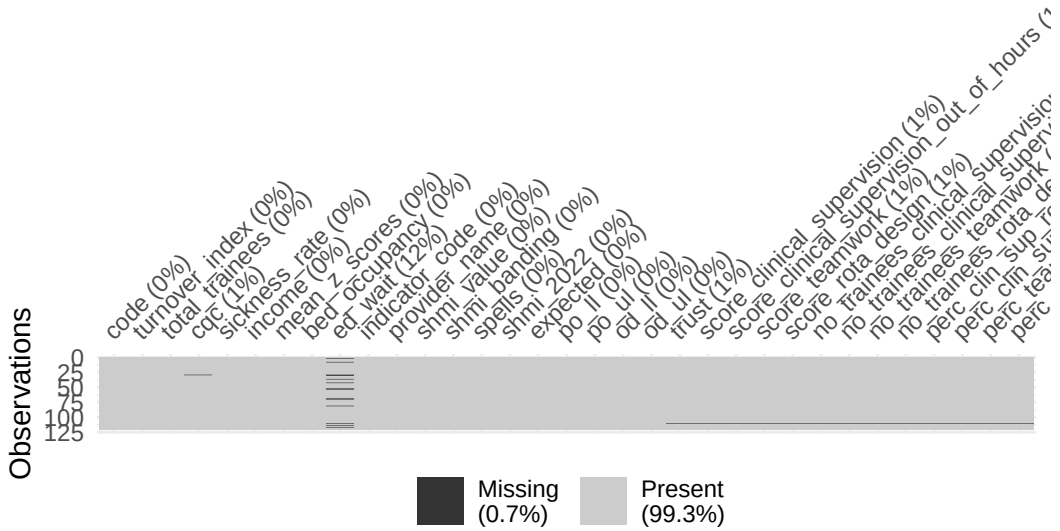
2 Methods

All datasets used are publically available. The National Training Survey was obtained from the GMC [@gmc-nts]. Data on ED waiting times and bed occupancy, were obtained from NHS England Statistical Work Areas [@nhsstatistics]. The number of junior doctors, sickness rate, stability index were obtained from NHS Digital [@nhs-sickness-absence-rates]. Financial status was obtained from NHS England Financial Accounting and Reporting [@nhsengland]. Adult inpatient survey data were obtained from NHS Surveys [@nhssurveys]. The CQC ratings were obtained via an API through the base URL (followed by the hospital trust code): `https://api.cqc.org.uk/public/v1`. The year of data collection was selected to be 2022, and if monthly data was available, April 2022 was used. However, with CQC ratings, the last given rating was used.

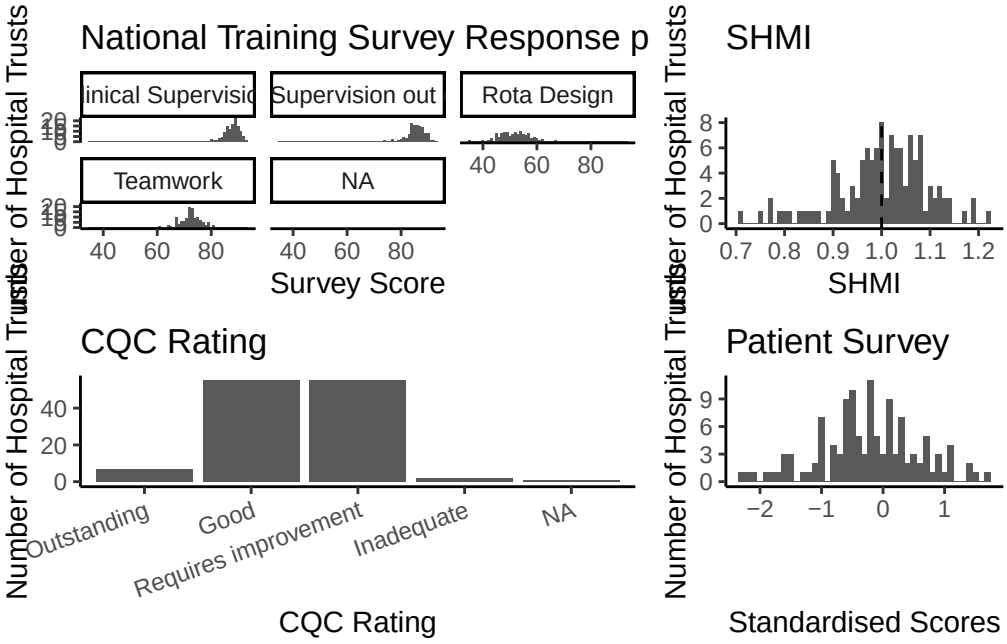
Survey responses were catagorised into four domains: 'clinical supervision,' 'clinical supervision outside regular hours,' 'teamwork,' and 'rota design. The domain scores were created by standardising the responses to questions that fell into these domains. Scores are out of 100, with 100 being the 'best' score. SHMI reflects mortality rates in hospitals by comparing the actual number of patient deaths to the expected number of deaths based on a pre-determined model. An SHMI score of 1 indicates that the actual number of patient deaths is the same as the expected number of patient deaths. The Care Quality Commission (CQC) rating is a standardised assessment of healthcare providers' quality and safety of care in England. It is measured through comprehensive inspections, evaluations, and assessments of various aspects of healthcare services, resulting in ratings ranging from "Outstanding" to "Inadequate" based on their performance in various domains, including safety, effectiveness, and responsiveness to patients' needs. The adult inpatient survey sampled inpatients aged 16 or over who were discharged from an NHS trust in England. The survey is divided in 11 main sections, which have a standardised score based on survey response rate. A mean of the standardised score for all these sections was taken as the overall survey score. National training survey scores, SHMI and patient survey scores are considered as continuous variables and CQC rating is considered as an ordinal variable.

2.1 Exploratory Data Analysis

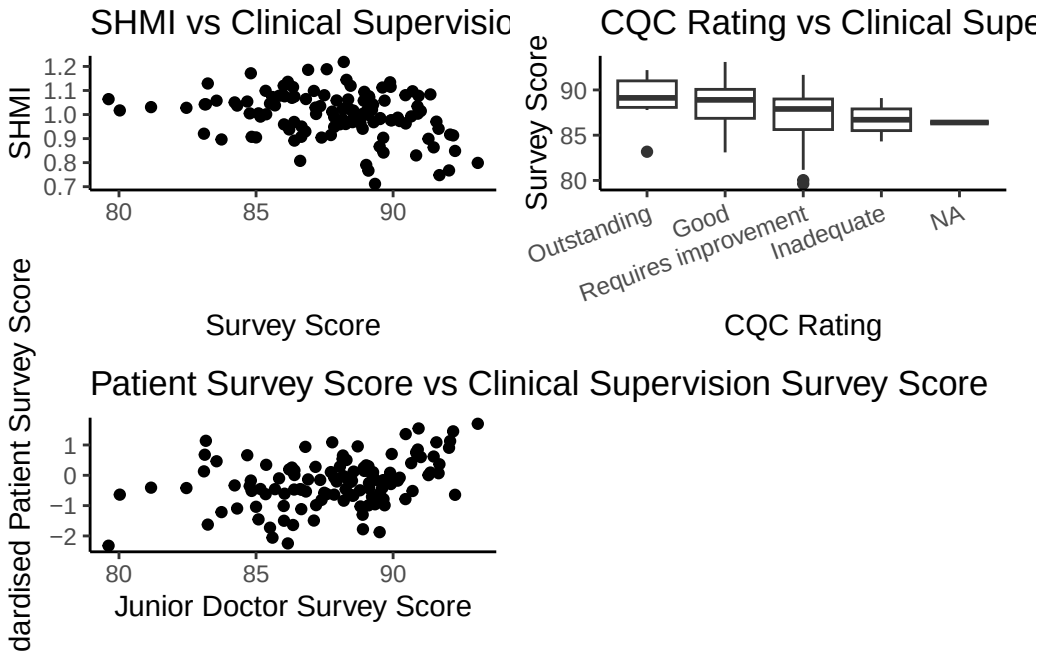
There were 120 foundation hospital trusts in England included. There are no missing data for SHMI or patient survey outcomes and there are no missing data for the NTS survey responses. There is one missing value for CQC rating, as the trust in question had not yet been rated. There is only missing data for ED waiting times, where 14 trusts do not have data. There did not appear to be an association between missingness and outcome variables, therefore the missing data were classified missing completely at random. A graphical representation of missingness is seen below. Given the low amount of missing data, I did not think it necessary to perform further analysis or imputation.



Histograms of survey scores for the four domains can be seen below. The responses are largely normally distributed, with rota design scored much worse than the other domains with a larger standard deviation. The histogram for SHMI is seen below. As expected, given this is a standardised score, the results are normally distributed.



The graphical association between the national training survey responses for the clincial supervision domain and each outcome measure are shown below. There appear to be a negative correlation between national training survey scores and SHMI and a positive correlation between national training survey scores and inpatient survey scores and CQC rating. The relationship between national training survey scores and both SHMI and inpatient survey scores appears to be non-linear. This will be explored further in the model building section of the results. The association for the other domains of the national training survey and outcome measures are not shown but follow a similar pattern.



Exploratory data analysis for other covariates was performed but is not shown. All covariates appear to be normally distributed with no outliers, with the exception of income, which has two outliers either side from the bulk of the observations, which are otherwise normally distributed. The hospital trusts from which these outliers came from were examined, but the data seem reliable and correct, and so it was decided to keep these outliers in the dataset. Sensitivity analysis was performed with and without the outliers (analysis not shown), with no significant change in results. The number of trainees was also not normally distributed, as the data exhibited a significant right skew.

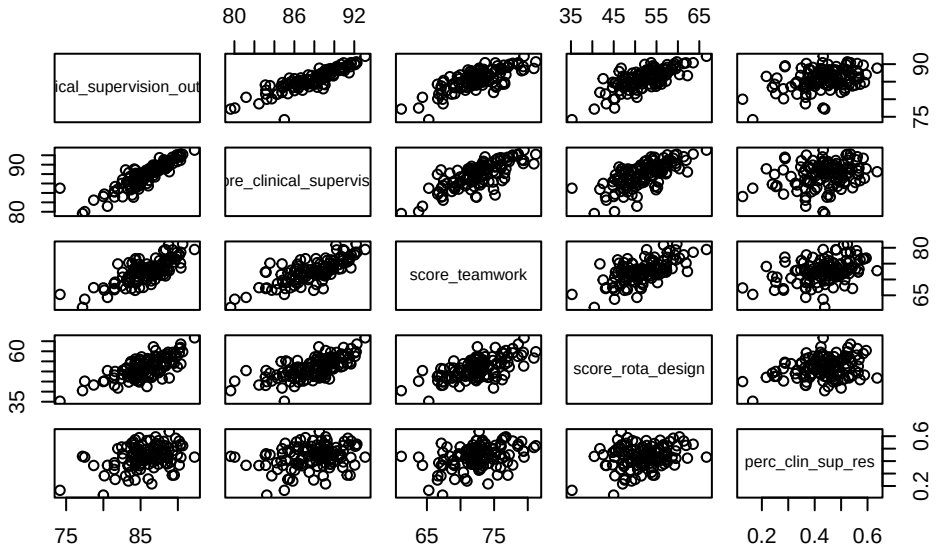
Of note, for the four domains of the national training survey there was a mean response rate of between 40.0% - 44.0%, with a minimum response rate of between 11.6% - 13.0% and a maximum response rate of between 57.7% - 61.8%. There was a median of 419 trainees per hospital trust, with a minimum of 139 and a maximum of 1829.

3 Results

In order to test the association between national training survey response and SHMI, CQC rating and inpatient survey score, I built two final linear regression models and an ordinal regression model. In the process of model selection, several models were built using flexible modelling, interaction terms, comparing full and reduced models and using regularisation techniques. These models were then compared using various model metrics, before three final models were selected.

3.1 Model Building & Selection - SHMI

For all the models that were constructed without regularisation techniques, only one of the four national survery domains were used as each domain was highly corrlated to each other, resulting in collinearity, The pairwise scatter plot illustrating this can be seen below. Clinical supervision was chosen to be included in the models.



The final regression model used a B-spline to flexibly model the relationship between clinical supervision survey response and SHMI. No interaction terms were seen to be significant, and therefore were not included in the final mode. Upon using the F-test, there was a significant difference between the linear model and the model including the B-spline, and there was also a significant difference between the full and reduced model. The metrics for all the models constructed can be seen in the table below. The final model was selected based on a low Root Mean Squared Error (RMSE), a high R-squared value and a low AIC (Akaike Information

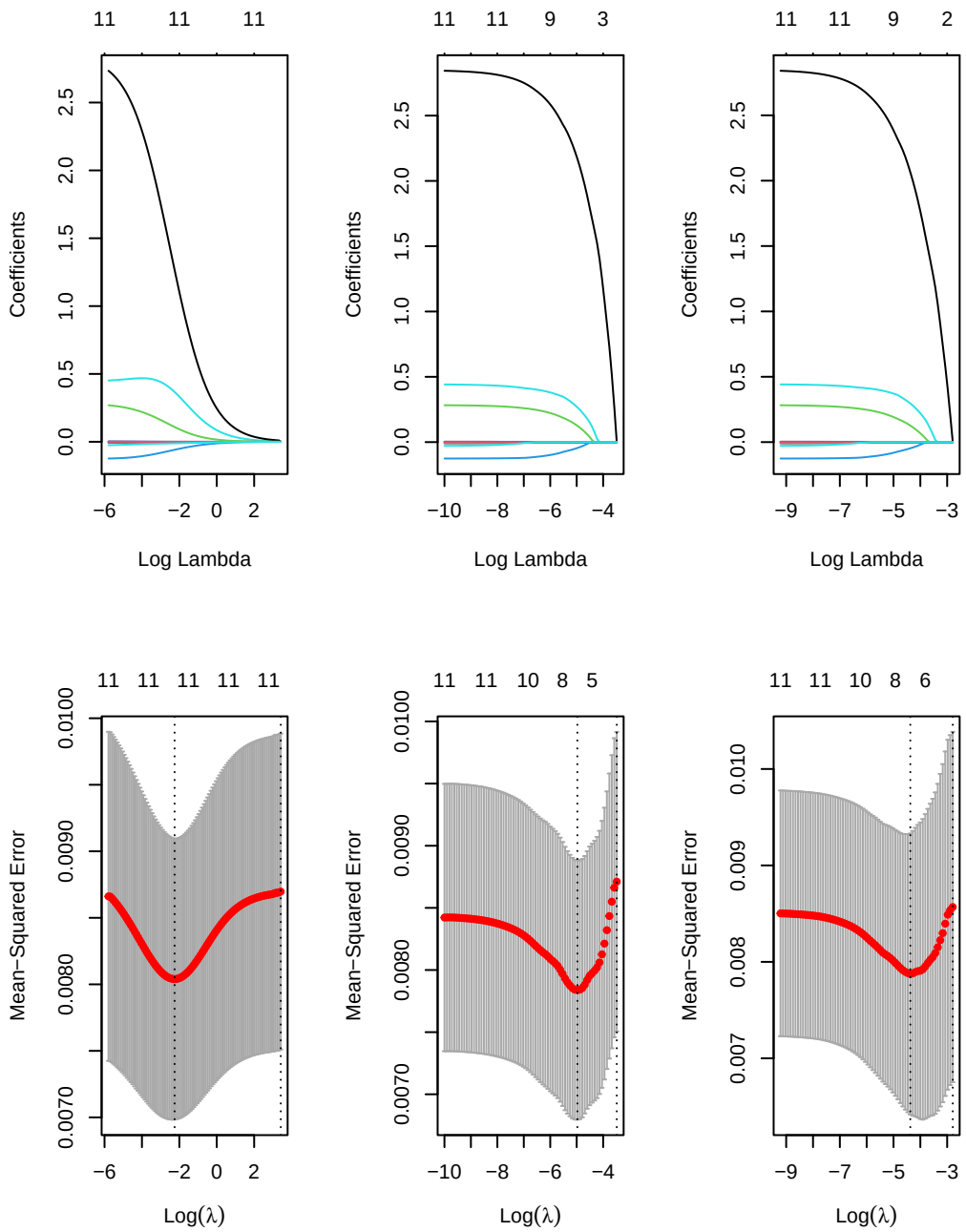
Model	RMSE	R2	AIC
Linear	0.080	0.225	-213.371
Quadratic	0.080	0.243	-213.914
B-spline	0.078	0.275	-216.479
Natural Cubic Spline	0.078	0.268	-215.387
GAM	0.078	NA	-215.278
Reduced	0.078	0.269	-219.632

term	estimate	std.error	statistic	p.value	significant
(Intercept)	0.597	0.426	1.402	0.164	N
bSpline(score_clinical_supervision, df = 3)1	-0.182	0.131	-1.386	0.169	N
bSpline(score_clinical_supervision, df = 3)2	0.106	0.080	1.333	0.186	N
bSpline(score_clinical_supervision, df = 3)3	-0.213	0.082	-2.609	0.011	Y
turnover_index	0.263	0.462	0.570	0.570	N
total_trainees	0.000	0.000	-0.351	0.726	N
sickness_rate	2.813	0.894	3.146	0.002	Y
income	0.000	0.000	0.242	0.809	N
bed_occupancy	0.210	0.157	1.339	0.184	N
ed_wait	-0.091	0.108	-0.849	0.398	N
perc_clin_sup_res	-0.075	0.095	-0.790	0.431	N

Criterion). Model diagnostics were performed (not shown), including checking for linearity, homoscedasticity, independence and normality. Multicollinearity was also assessed using variance inflation factor (VIF). All model assumptions were met. There were no significant outliers or observations with high influence.

The output of the final mode is seen below. We can see that the the national training survey score for how junior doctors feel about clinical supervision is significantly associated with SHMI. The sickness rate is also significantly associated with SHMI. The estimate is negative, indicating that as junior doctors rate their experience of clinical supervision more highly, the SHMI decreases. Conversely, as sickness rates increase, SHMI also tends to increase.

To check for overfitting, I performed regularisation using LASSO, Ridge and Elastic net. The results are below. By using penalities to prevent overfitting, the LASSO method kept only the clinical supervision and teamwork domains in the national training survey, stability index and sickness rate. This is in agreement with the models described above.


11 11 11
11 11 9 3
11 11 9 2
Coefficients
Log Lambda
Mean−Squared Error
Log(λ)
11 11 11 11 11
11 11 10 8 5
11 11 10 8 6

	ridge	lasso	elastic
(Intercept)	1.003	1.032	1.019
score_clinical_supervision	-0.003	-0.003	-0.003
score_clinical_supervision_out_of_hours	-0.001	0.000	0.000
score_teamwork	-0.002	-0.003	-0.003
score_rota_design	0.000	0.000	0.000
turnover_index	0.367	0.266	0.296
total_trainees	0.000	0.000	0.000
sickness_rate	1.266	2.170	2.065
income	0.000	0.000	0.000
bed_occupancy	0.109	0.125	0.127
ed_wait	-0.059	-0.045	-0.047
perc_clin_sup_res	-0.011	0.000	0.000

Model	RMSE	R2	AIC
Linear	0.704	0.243	246.466
Quadratic	0.682	0.291	241.607
B-spline	0.637	0.381	229.244
Natural Cubic Spline	0.649	0.358	233.120
GAM	0.645	NA	231.898
Reduced	0.664	0.327	234.058
Interaction	0.635	0.385	230.460

3.2 Model Building & Selection - Inpatient Survey Scores

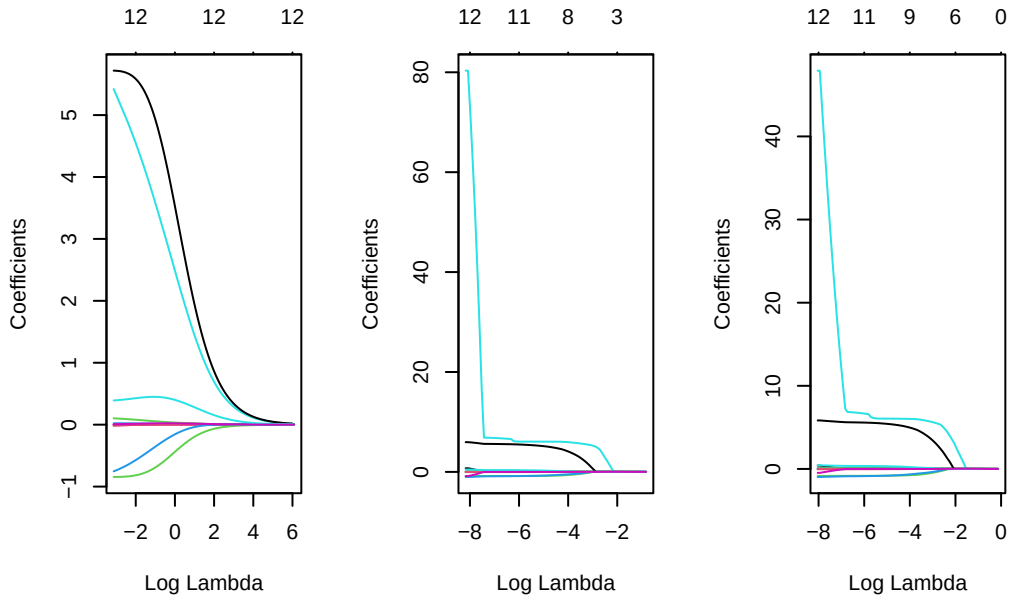
Next, I built models to explore the association between national training survey clinical supervision scores and inpatient survey scores. As before, flexible modelling and interaction terms were explored. The best model fit was provided when clinical supervision scores were modelled with a natural cubic spline. Interestingly, an interaction effect between clinical supervision scores and the stability index was seen. The full model was shown to be significantly different compared to the reduced model. A table of model metrics can be seen below. Model diganostics were performed to check that all model assumptions were met (not shown). There were no deviations from model assumptions.

The final model output is shown below. Scores for the national training survey clinical supervision domain were positively associated with inpatient survey scores. The stability index (seen here as turnover index) was significantly positively associated with inpatient survey scores, indicating that when more staff remain in their position, patient survey scores tend to increase. There is a significant interaction term between scores for the national training survey clinical supervision domain and stability index. The estimate is negative, indicating that as scores for the national training survey clinical supervision domain increase, the effect of the stability index on inpatient survey scores is decreased, and vice versa. Finally, there is also a significant

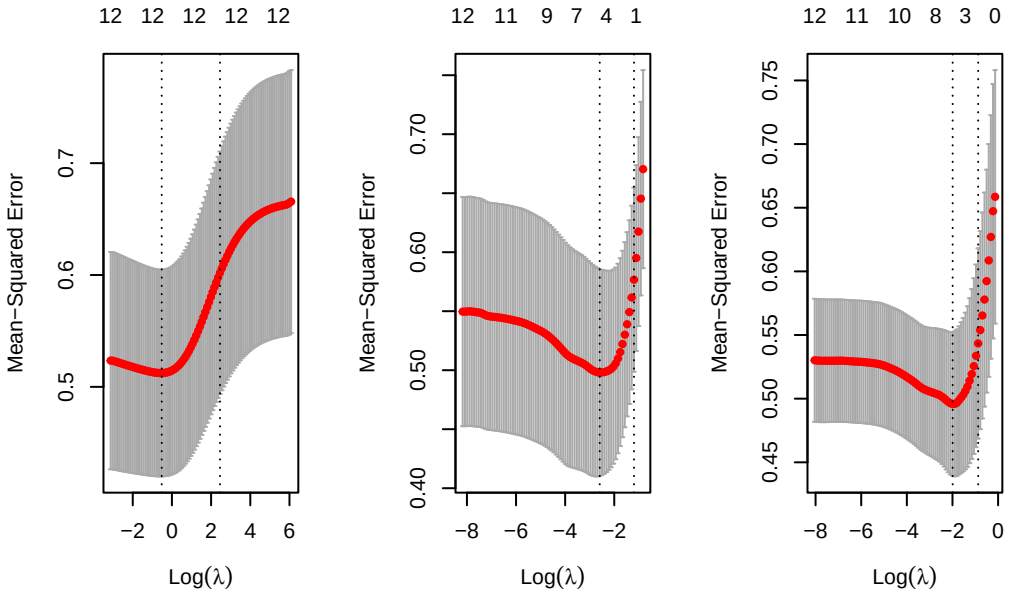
term	estimate	std.error	statistic	p.value	significant
(Intercept)	-31.006	10.690	-2.900	0.005	Y
ns(score_clinical_supervision, df = 3)1	19.688	9.592	2.052	0.043	Y
ns(score_clinical_supervision, df = 3)2	43.022	19.913	2.161	0.033	Y
ns(score_clinical_supervision, df = 3)3	27.236	12.224	2.228	0.028	Y
total_trainees	0.000	0.000	0.802	0.425	N
turnover_index	257.515	120.298	2.141	0.035	Y
sickness_rate	6.189	7.344	0.843	0.401	N
income	0.000	0.000	0.573	0.568	N
bed_occupancy	-0.138	1.278	-0.108	0.914	N
ed_wait	-1.187	0.914	-1.299	0.197	N
perc_clin_sup_res	1.846	0.779	2.370	0.020	Y
score_clinical_supervision:turnover_index	-2.798	1.361	-2.055	0.043	Y

positive association between the percentage of junior doctors who responded to the survey and inpatient survey scores.

As above, regularisation methods were used to check for overfitting. The results are below. The LASSO model included the clinical supervision, teamwork and rota design domains in the final model, along with stabiliy index and the interaction term between clinical supervision and stability index. Interestingly, the percentage of junior doctors who responded to the survey was not included, possibly due to collinearity.



	ridge	lasso	elastic
(Intercept)	-10.556	-10.641	-10.019
score_clinical_supervision	0.013	0.000	0.000
score_clinical_supervision_out_of_hours	0.014	0.000	0.000
score_teamwork	0.044	0.088	0.076
score_rota_design	0.015	0.009	0.011
turnover_index	3.066	3.599	2.710
total_trainees	0.000	0.000	0.000
sickness_rate	4.335	0.000	0.000
income	0.000	0.000	0.000
bed_occupancy	-0.580	0.000	0.000
ed_wait	-0.252	0.000	0.000
perc_clin_sup_res	0.437	0.000	0.000
score_clinical_supervision:turnover_index	0.022	0.006	0.018



3.3 Model Building & Selection - CQC Rating

The final outcome measure I assessed was CQC rating, which I treated as an ordinal variable. The full model was significantly different to the reduced model, and so the full model was chosen. Model diagnostic were performed (not shown) and all model assumptions were met.

The model output is shown below. A higher CQC rating is significantly associated with national training survey scores in the clinical supervision domain, a higher hospital trust income and the percentage of junior doctors who responded to the survey. The bed utilisation rate tended towards significance, but did not meet the 0.05 P-value cut-off.

```
Call:
vglm(formula = cqc ~ score_clinical_supervision + turnover_index +
    total_trainees + sickness_rate + income + bed_occupancy +
    ed_wait + perc_clin_sup_res, family = cumulative(parallel = TRUE,
    reverse = TRUE), data = data)

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept):1               1.842e+01  1.250e+01   1.474  0.14046
(Intercept):2               1.501e+01  1.245e+01   1.205  0.22805
(Intercept):3               1.039e+01  1.241e+01   0.837  0.40268
score_clinical_supervision -2.424e-01  9.077e-02  -2.671  0.00756 **
turnover_index              5.785e+00  1.125e+01   0.514  0.60715
total_trainees              9.053e-04  6.843e-04   1.323  0.18584
sickness_rate              -1.235e+01  2.220e+01  -0.556  0.57809
income                     -1.712e-05  6.703e-06  -2.554  0.01064 *
bed_occupancy               7.132e+00  3.829e+00   1.863  0.06251 .
ed_wait                    -3.284e+00  2.698e+00  -1.217  0.22357
perc_clin_sup_res          -5.599e+00  2.488e+00  -2.250  0.02444 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Names of linear predictors: logitlink(P[Y>=2]), logitlink(P[Y>=3]),
logitlink(P[Y>=4])

Residual deviance: 172.3349 on 307 degrees of freedom

Log-likelihood: -86.1674 on 307 degrees of freedom

Number of Fisher scoring iterations: 6

No Hauck-Donner effect found in any of the estimates


Exponentiated coefficients:
score_clinical_supervision             turnover_index
              7.847070e-01               3.253148e+02
```

```
total_trainees              sickness_rate
  1.000906e+00               4.345622e-06
        income              bed_occupancy
  9.999829e-01               1.251370e+03
       ed_wait          perc_clin_sup_res
  3.748300e-02               3.701897e-03
```

4 Conclusion

This study investigated the association between junior doctors' evaluations of their working environment in foundation hospital trusts in England and three key hospital performance metrics: Summary Hospital-level Mortality Indicator (SHMI), inpatient survey scores, and Care Quality Commission (CQC) ratings. The analysis revealed significant associations between junior doctors' scores in categories such as clinical supervision, clinical supervision out of hours, teamwork, rota design, and the aforementioned hospital performance metrics. Additionally, the sickness rate showed an association with SHMI, while the stability index and the percentage of junior doctors responding to the national training survey were linked to inpatient survey scores. Hospital trust income and the percentage of junior doctors responding to the survey were associated with CQC ratings. This study produced three robust regression models that accurately explained a large proportion of the variance in SHMI, inpatient survey score and CQC rating between trusts. The linear regression models had R-squared values of 0.28 and 0.36, indicating a good fit to the data. Overfitting was avoided by using regularisation techniques.

The results of this study are in line with previous research demonstrating an association between doctor staff numbers and hospital mortality, and doctor burnout and patient safety, as well as one study demonstrating a link between hospital income and hospital performance measures [@harvey2021association] [@hodkinson2022associations] [@nagendran2019financial].

However, several limitations should be considered. The study encountered challenges related to the low response rate of junior doctors to the national training survey, potentially introducing bias into the results. Those that did not respond may have done so because they felt stressed or too busy, and may have been more likely to be in hospital trusts whether other junior doctors rated their training programme poorly. Furthermore, there is uncertainty as to whether the national training survey, CQC ratings, inpatient survey and SHMI truely capture unbiased reflections of junior doctors' experience and hospital performance.

While these findings are intriguing, determining the direction of causality is challenging. It is plausible that junior doctors' negative evaluations of their training programmes contribute to poor hospital outcomes, just as it is plausible that poor hospital outcomes influence junior doctors' perceptions of their training programs. If junior doctors' perception of their training programme influences hospital outcomes, there is a clear imperitive to improve the training

programmes, not only to improve the quality of lives of junior doctors, but also to improve patient outcomes.

The ongoing influence of the COVID-19 pandemic on the results remains a subject of exploration. To ensure the robustness of the findings, it would be valuable to validate them by analysing pre-COVID data. Conducting time series analyses could offer insights into the temporal fluctuations in hospital outcomes and whether changes in junior doctor responses precede or follow changes in hospital performance. Understanding the timing of events could contribute to a better grasp of causality. Future research should also explore in further depths the experiences of junior doctors and how this may impact, or be impacted by, hospital performance, via the use of qualitative research. Research should also be conducted on how hospital policies can impact and improve the experience of junior doctors.

In conclusion, this study underscores the significance of junior doctors' evaluations in foundation hospital trusts in England and their association with critical hospital performance metrics. Junior doctors represent the canary in the goldmine, and greater attention should be given to their experiences inside the hospital.

4.0.1 References

@online{nhsengland, author = {NHS England}, title = {NHS providers: trust accounts consolidation (TAC) data publications}, year = {2022}, url = {https://www.england.nhs.uk/financial-accounting-and-reporting/nhs-providers-tac-data-publications/} }
@online{nhsstatistics, author = {NHS England}, title = {Statistical Work Areas - NHS England}, year = {2022}, url = {https://www.england.nhs.uk/statistics/statistical-work-areas/} }
@online{nhssurveys, author = {NHS Surveys}, title = {NHS Surveys}, year = {2022}, url = {https://nhssurveys.org/surveys/} }
@online{nhs-sickness-absence-rates, author = {NHS Digital}, title = {NHS Sickness Absence Rates (July 2023) - Provisional Statistics}, year = {2022}, url = {https://digital.nhs.uk/data-and-information/publications/statistical/nhs-sickness-absence-rates/july-2023-provisional-statistics} }
@online{nhs-workforce-statistics, author = {NHS Digital}, title = {NHS Workforce Statistics (March 2023)}, year = {2022}, url = {https://digital.nhs.uk/data-and-information/publications/statistical/nhs-workforce-statistics/march-2023} }
@online{gmc-nts, author = {General Medical Council}, title = {National Training Surveys}, year = {Year of Publication or Access}, url = {https://www.gmc-uk.org/education/how-we-quality-assure-medical-education-and-training/evidence-data-and-intelligence/national-training-surveys} }
@article{hodkinson2022associations, author = {Hodkinson, A. and Zhou, A. and Johnson, J. and Geraghty, K. and Riley, R. and Zhou, A. et al.}, title = {Associations of physician

burnout with career engagement and quality of patient care: systematic review and meta-analysis}, journal = {BMJ}, year = {2022}, volume = {378}, pages = {e070442}, doi = {10.1136/bmj-2022-070442} }
@article{harvey2021association, title = {The association between physician staff numbers and mortality in English hospitals}, author = {Harvey, Philip R. and Trudgill, Nigel J.}, journal = {EClinicalMedicine}, volume = {32}, pages = {100709}, year = {2021}, month = {February}, doi = {10.1016/j.eclinm.2020.100709}, url = {https://doi.org/10.1016/j.eclinm.2020.100709}, note = {Open Access}, published = {January 06, 2021}, }
@article{nagendran2019financial, title = {Financial performance of English NHS trusts and variation in clinical outcomes: a longitudinal observational study}, author = {Nagendran, Myura and Kiew, Gavin and Raine, Rosalind and Atun, Rifat and Maruthappu, Mahiben}, journal = {BMJ Open}, volume = {9}, number = {1}, pages = {e021854}, year = {2019}, month = {Jan}, day = {28}, doi = {10.1136/bmjopen-2018-021854}, pmid = {30696667}, pmcid = {PMC6352807}, }

5 Appendix (code)

```
library(tidyverse)
library(stringr)
library(knitr)
library(kableExtra)
library(glmnet)
library(caret)
library(patchwork)
library(lme4)
library(naniar)
library(lmerTest)
library(httr)
library(jsonlite)
library(stringr)
library(stringdist)
library(fuzzyjoin)
library(readxl)
library(VGAM)
library(splines)
library(splines2)
library(gam)
library(patchwork)
library(broom)
library(brant)
```

```r
library(car)

#As the API throttles when I repeated try to render the document I saved the final dataset
#I've copied some of the code from the codeblocks to make sure the rest of the code runs

data <- read_csv("all_trust_data_combined.csv")
data$cqc <- ordered(data$cqc, levels = c("Outstanding", "Good", "Requires improvement", "I

load("All_data_workspace.RData")
load("Individual Question Analysis Data.RData")
nts_per_trust <- read_excel("nts_questions_trust.xlsx", sheet = "8. Trust Scores")
colnames(nts_per_trust) <- c("year", "trust", "domain", "score", "no_trainees")

#Filter for 2022
nts_per_trust_2022 <- nts_per_trust %>%
  filter(year == 2022)

shmi_2022 <- SHMI_2022 %>%
  filter(Year == 2022)
colnames(shmi_2022) <- tolower(colnames(shmi_2022))
colnames(shmi_2022)[4] <- "shmi_value"

#Join SHMI and NTS datasets
shmi_nts_2022 <- stringdist_left_join(shmi_2022, nts_per_trust_2022,
                                      by = c("provider_name" = "trust"),
                                      method = "jw",
                                      max_dist = 0.01,
                                      ignore_case = TRUE)

#convert into numeric
shmi_nts_2022$score <- as.numeric(shmi_nts_2022$score)

shmi_nts_2022_wider <- pivot_wider(shmi_nts_2022, names_from = domain, values_from = c(sco
colnames(shmi_nts_2022_wider) <- tolower(colnames(shmi_nts_2022_wider))
colnames(shmi_nts_2022_wider) <- gsub(" ", "_", colnames(shmi_nts_2022_wider))

#Load NTS data
setwd("~/Research/METRO/NTS")
load("All_data_workspace.RData")
load("Individual Question Analysis Data.RData")
nts_per_trust <- read_excel("nts_questions_trust.xlsx", sheet = "8. Trust Scores")
```

```r
colnames(nts_per_trust) <- c("year", "trust", "domain", "score", "no_trainees")

#load SHMI data
setwd("~/Research/METRO/SHMI")
shmi_files <- list.files(pattern = "\\.csv$")

#load ED waiting times
setwd("~/Research/METRO/NTS")
ed_waiting <- read_excel("ae_waitingtimes.xls", sheet = 2, skip = 15)
ed_waiting <- ed_waiting[3:nrow(ed_waiting),]
colnames(ed_waiting) <- tolower(colnames(ed_waiting))
colnames(ed_waiting) <- gsub(" ", "_", colnames(ed_waiting))
colnames(ed_waiting) <- gsub("\\(", "", colnames(ed_waiting))
colnames(ed_waiting) <- gsub("\\)", "", colnames(ed_waiting))
ed_waiting <- ed_waiting %>%
  select(code, name, percentage_in_4_hours_or_less_all)
colnames(ed_waiting) <- c("provider_code", "trust", "ed_wait")

#load bed occupancy
beds_occupied <- read_excel("beds_occupied.xlsx", sheet = 1, skip = 14)
beds_occupied <- beds_occupied[3:nrow(beds_occupied),]
colnames(beds_occupied) <- tolower(colnames(beds_occupied))
colnames(beds_occupied) <- gsub(" ", "_", colnames(beds_occupied))
beds_occupied <- beds_occupied %>%
  select(18, org_code, org_name)
colnames(beds_occupied) <- c("bed_occupancy", "provider_code", "trust")


#load IP survey data
survey <- read_excel("ip_survey_2022.xlsx", sheet = 4)
z_scores <- grep("final", names(survey), value = TRUE)
z_scores <- grep("S", z_scores, value = TRUE)
columns <- c("TrustCode", "trustname", z_scores)
survey <- survey[, columns]
survey <- survey %>%
  mutate(mean_z_scores = rowMeans(select(., all_of(z_scores)), na.rm = TRUE))


#load finanical status
finance <- read_excel("financial_status_2022.xlsx", sheet = 3)
finance_codes <- read_excel("financial_status_2022.xlsx", sheet = 2)
```

```r
finance_codes <- finance_codes %>%
  select(`Full name of Provider`, `NHS code`)
finance <- finance %>%
  filter(MainCode == "A02CY01" & SubCode == "SOC0190") %>%
  select(`Organisation Name`, `Value number`)
finance <- left_join(finance, finance_codes, by = c("Organisation Name" = "Full name of Pr
colnames(finance) <- c("trust", "income", "code")
finance[finance$trust == "Guy's and St Thomas' NHS Foundation Trust", 3] <- "RJ1"
finance[finance$trust == "King's College Hospital NHS Foundation Trust", 3] <-"RJZ"
finance[finance$trust == "Birmingham Women's and Children's NHS Foundation Trust", 3] <-"R

finance_2 <- read_excel("financial_status_2_2022.xlsx", sheet = 3)
finance_2_codes <- read_excel("financial_status_2_2022.xlsx", sheet = 2)
finance_2_codes <- finance_2_codes %>%
  select(`Full name of Provider`, `NHS code`)
finance_2 <- finance_2 %>%
  filter(MainCode == "A02CY01" & SubCode == "SOC0190") %>%
  select(`Organisation Name`, `Value number`)
finance_2 <- left_join(finance_2, finance_2_codes, by = c("Organisation Name" = "Full name
colnames(finance_2) <- c("trust", "income", "code")
finance_2[finance_2$trust == "St Helens And Knowsley Teaching Hospitals NHS Trust", 3] <-
finance_2[finance_2$trust == "West Hertfordshire Teaching Hospitals NHS Trust", 3] <- "RWG

finance <- rbind(finance, finance_2)


#load sickness rates
sickness <- read_excel("sickness_2022.xlsx", sheet = 6, skip = 9)
sickness <- sickness %>%
  select(3, 4, 113)
colnames(sickness) <- c("trust", "code", "sickness_rate")

#get CQC dat from API
codes <- survey$TrustCode

cqc <- data.frame()
for (code in codes) {
  api_url <- paste0("https://api.cqc.org.uk/public/v1/providers/", code)
  response <- GET(api_url)
  data <- fromJSON(rawToChar(response$content))
```

```
  # Check for 'combinedQualityRating' in 'useOfResources' section
  if (!is.null(data$currentRatings$overall$useOfResources$combinedQualityRating)) {
    quality_rating <- data$currentRatings$overall$useOfResources$combinedQualityRating
  } else {
    # If not found, check in 'overall' section
    quality_rating <- data$currentRatings$overall$rating
  }

  # Store the result in 'cqc' dataframe
  cqc <- rbind(cqc, data.frame(
    Provider_Code = code,
    Quality_Rating = quality_rating
  ))
  Sys.sleep(2)
}


colnames(cqc) <- c("code", "cqc")
# Replace "No published rating" with NA in the "CQC_Rating" column
cqc$cqc <- ifelse(cqc$cqc == "No published rating", NA, cqc$cqc)



#number of trainees (can also look at number of other doctors/HCWs)
no_trainees <- read_excel("number_of_trainees_2022.xlsx", sheet = 3, skip = 6)
no_trainees <- no_trainees[5:nrow(no_trainees),]
colnames(no_trainees)[3:4] <- c("trust", "code")
colnames(no_trainees) <- gsub(" ", "_", colnames(no_trainees))
no_trainees <- no_trainees %>%
  select(trust, code, Total, Specialty_Doctor, Specialty_Registrar, Core_Training, Foundat
  mutate(all_trainees = rowSums(select(., Specialty_Doctor:Foundation_Doctor_Year_1), na.r
  select(trust, code, Total, all_trainees)
colnames(no_trainees)[3:4] <- c("total_staff", "total_trainees")
no_trainees <- na.omit(no_trainees)

fulltime_trainees <- read_excel("number_of_trainees_2022.xlsx", sheet = 4, skip = 6)
fulltime_trainees <- fulltime_trainees[5:nrow(fulltime_trainees),]
colnames(fulltime_trainees)[3:4] <- c("trust", "code")
colnames(fulltime_trainees) <- gsub(" ", "_", colnames(fulltime_trainees))
fulltime_trainees <- fulltime_trainees %>%
  select(trust, code, Total, Specialty_Doctor, Specialty_Registrar, Core_Training, Foundat
```

```
  mutate(all_trainees = rowSums(select(., Specialty_Doctor:Foundation_Doctor_Year_1), na.r
  select(trust, code, Total, all_trainees)
colnames(fulltime_trainees)[3:4] <- c("fulltime_staff", "fulltime_trainees")
fulltime_trainees <- na.omit(fulltime_trainees)

number_of_trainees <- left_join(no_trainees, fulltime_trainees, by = "code")
number_of_trainees <- number_of_trainees %>%
  mutate(perc_fullime_trainees = fulltime_trainees / total_trainees)


#load turnover
turnover <- read_excel("turnover_2022.xlsx", sheet = 1, skip = 7)
colnames(turnover) <- gsub(" ", "_", colnames(turnover))
turnover <- turnover %>%
  filter(Staff_group == "All staff groups") %>%
  select(3, 4, 12)
colnames(turnover) <- c("trust", "code", "turnover_index")

#Remove CCGs, PCTs etc.
nts_per_trust <- nts_per_trust %>%
  filter(!grepl("CCG", trust, ignore.case = TRUE) &
           !grepl("PCT", trust, ignore.case = TRUE))


#Filter for 2022
nts_per_trust_2022 <- nts_per_trust %>%
  filter(year == 2022)

shmi_2022 <- SHMI_2022 %>%
  filter(Year == 2022)
colnames(shmi_2022) <- tolower(colnames(shmi_2022))
colnames(shmi_2022)[4] <- "shmi_value"


#Renames some trusts in the NTS dataset so they join properly

nts_per_trust_2022 <- nts_per_trust_2022 %>%
  mutate(trust = ifelse(trust == "South Warwickshire NHS Foundation Trust",
                                "SOUTH WARWICKSHIRE UNIVERSITY NHS FOUNDATION TRUST",
                                trust),
         trust = ifelse(trust == "Royal Devon and Exeter NHS Foundation Trust",
```

```
                                "ROYAL DEVON UNIVERSITY HEALTHCARE NHS FOUNDATION TRUST",
                                trust),
         trust = ifelse(trust == "Mid Yorkshire Hospitals NHS Trust",
                                "MID YORKSHIRE TEACHING NHS TRUST",
                                trust),
         trust = ifelse(trust == "Homerton University Hospital NHS Foundation Trust",
                                "HOMERTON HEALTHCARE NHS FOUNDATION TRUST",
                                trust))



#Join SHMI and NTS datasets
shmi_nts_2022 <- stringdist_left_join(shmi_2022, nts_per_trust_2022,
                                      by = c("provider_name" = "trust"),
                                      method = "jw",
                                      max_dist = 0.01,
                                      ignore_case = TRUE)

#convert into numeric
shmi_nts_2022$score <- as.numeric(shmi_nts_2022$score)

shmi_nts_2022_wider <- pivot_wider(shmi_nts_2022, names_from = domain, values_from = c(sco
colnames(shmi_nts_2022_wider) <- tolower(colnames(shmi_nts_2022_wider))
colnames(shmi_nts_2022_wider) <- gsub(" ", "_", colnames(shmi_nts_2022_wider))
#combine rest of the datasets
shmi_nts_2022_wider <- shmi_nts_2022_wider %>%
  select(-year.x, -year.y)
#ed waiting times
shmi_nts_ed_2022 <- ed_waiting %>%
  select(provider_code, ed_wait) %>%
  right_join(shmi_nts_2022_wider, ed_waiting,
                                  by= "provider_code")
shmi_nts_ed_2022$ed_wait <- as.numeric(shmi_nts_ed_2022$ed_wait)


#bed occupancy
shmi_nts_ed_beds_2022 <- beds_occupied %>%
  select(provider_code, bed_occupancy) %>%
  right_join(shmi_nts_ed_2022, beds_occupied,
                                  by= "provider_code")
shmi_nts_ed_beds_2022$bed_occupancy <- as.numeric(shmi_nts_ed_beds_2022$bed_occupancy)
```

```r
#IP survey
shmi_nts_ed_beds_survey<- survey  %>%
  select(TrustCode, mean_z_scores) %>%
  right_join(shmi_nts_ed_beds_2022, by = c("TrustCode" = "provider_code"))


#finanical status
shmi_nts_ed_beds_survey_finance <- finance %>%
  select(code, income) %>%
  right_join(shmi_nts_ed_beds_survey, by = c("code" = "TrustCode"))

#sickness
shmi_nts_ed_beds_survey_finance_sickness <- sickness %>%
  select(code, sickness_rate) %>%
  right_join(shmi_nts_ed_beds_survey_finance, by = "code")

#cqc
shmi_nts_ed_beds_survey_finance_sickness_cqc <- cqc %>%
  right_join(shmi_nts_ed_beds_survey_finance_sickness, by = "code")

#number of trainees
shmi_nts_ed_beds_survey_finance_sickness_cqc_trainees <- no_trainees %>%
  select(code, total_trainees) %>%
  right_join(shmi_nts_ed_beds_survey_finance_sickness_cqc, by = "code")

#turnover
shmi_nts_ed_beds_survey_finance_sickness_cqc_trainees_turnover <- turnover %>%
  select(code, turnover_index) %>%
  right_join(shmi_nts_ed_beds_survey_finance_sickness_cqc_trainees, by = "code")

#relabel dataframe to 'data'
data <- shmi_nts_ed_beds_survey_finance_sickness_cqc_trainees_turnover
data <- distinct(data)

#classify cqc into ordinal data
data$cqc <- factor(data$cqc, levels = c("Outstanding", "Good", "Requires improvement", "In
data$cqc <- ordered(data$cqc, levels = c("Outstanding", "Good", "Requires improvement", "I

#ensure other columns are the correct structure
data$sickness_rate <- as.numeric(data$sickness_rate)
data$no_trainees_clinical_supervision <- as.numeric(data$no_trainees_clinical_supervision)
```

```r
data$no_trainees_clinical_supervision_out_of_hours <- as.numeric(data$no_trainees_clinical
data$no_trainees_teamwork <- as.numeric(data$no_trainees_teamwork)
data$no_trainees_rota_design <- as.numeric(data$no_trainees_rota_design)

#calculate percentage of trainees who responsed
data$perc_clin_sup_res <- data$no_trainees_clinical_supervision / data$total_trainees
data$perc_clin_sup_ooh_res <- data$no_trainees_clinical_supervision_out_of_hours / data$to
data$perc_team_res <- data$no_trainees_teamwork / data$total_trainees
data$perc_rota_res <- data$no_trainees_rota_design / data$total_trainees

data <- data %>%
  select(-score_na,-no_trainees_na)

write.csv(data, "all_trust_data_combined.csv", row.names = FALSE)

#missingness of outcomes
shmi_missingness <- sum(complete.cases(data$shmi_value)) / length(data$shmi_value)
cqc_missingness <- sum(complete.cases(data$cqc)) / length(data$cqc)
survey_missingness <- sum(complete.cases(data$mean_z_scores)) / length(data$mean_z_scores)

#missingness of NTS data
clin_sup <- sum(complete.cases(data$score_clinical_supervision)) / length(data$score_clini
clin_sup_ooh <- sum(complete.cases(data$score_clinical_supervision_out_of_hours)) / length
teamwork <- sum(complete.cases(data$score_teamwork)) / length(data$score_teamwork)
rota_design <- sum(complete.cases(data$score_rota_design)) / length(data$score_rota_design

#complete cases
complete_cases <- sum(complete.cases(data)) / nrow(data)


#graphical display of missing ed_wait data and shmi/patient survey
ed_shmo_missing <- ggplot(data,
       aes(x = ed_wait,
           y = shmi_value)) +
 geom_miss_point()

ed_survey_missing <- ggplot(data,
       aes(x = ed_wait,
           y = mean_z_scores)) +
 geom_miss_point()
```

```r
#Missing data in whole dataset
vis_miss(data)


#Descriptive analysis of NTS data
nts_hist <- ggplot(shmi_nts_2022, aes(score)) +
  geom_histogram(binwidth = 1) +
  facet_wrap(~domain) +
  theme_classic() +
  labs(x = "Survey Score",  y = "Number of Hospital Trusts", title = "National Training Su

summary_clin_sup <- summary(data$score_clinical_supervision)
summary_clin_sup_ooh <- summary(data$score_clinical_supervision_out_of_hours)
summary_teamwork <- summary(data$score_teamwork)
summary_rota <- summary(data$score_rota_design)
#Descriptive analysis of SHMI data
#Histogram of shmi
shmi_hist <- ggplot(data, aes(shmi_value)) +
  geom_histogram(binwidth = 0.01) +
  geom_vline(xintercept = 1, linetype = "dashed") +
  theme_classic() +
  labs(x = "SHMI", y = "Number of Hospital Trusts", title = "SHMI")


#Number of trusts in each band
shmi_banding_barplot <- ggplot(shmi_nts_2022_wider, aes(x = shmi_banding)) +
  geom_bar() +
  ggtitle("Number of trusts in each band  \n 1 = lower than expected number of deaths, 2 =
  theme_classic()



#scatterplot
shmi_clin_sup <- ggplot(data, aes(x = score_clinical_supervision, y = shmi_value)) +
  geom_point() +
  theme_classic() +
  labs(x = "Survey Score", y = "SHMI", title = "SHMI vs Clinical Supervision Survey Score"

shmi_clin_sup_ooh <- ggplot(data, aes(x = score_clinical_supervision_out_of_hours, y = shm
  geom_point()
```

```r
shmi_teamwork <- ggplot(data, aes(x = score_teamwork, y = shmi_value)) +
  geom_point()

shmi_rota_design <- ggplot(data, aes(x = score_rota_design, y = shmi_value)) +
  geom_point()

summary_shmi <- summary(data$shmi_value)
#Descriptive analysis of CQC data
cqc_barplot <- ggplot(data, aes(x = cqc)) +
  geom_bar() +
  theme_classic() +
  labs(x = "CQC Rating", y = "Number of Hospital Trusts", title = "CQC Rating") +
  theme(axis.text.x = element_text(angle = 20, hjust = 1))

#boxplots
cqc_clin_sup <- ggplot(data, aes(x = cqc, y =  score_clinical_supervision )) +
  geom_boxplot() +
  theme_classic() +
  labs(y = "Survey Score", x = "CQC Rating", title = "CQC Rating vs Clinical Supervision S
  theme(axis.text.x = element_text(angle = 20, hjust = 1))

cqc_clin_sup_ooh <- ggplot(data, aes(x = cqc, y =  score_clinical_supervision_out_of_hours
  geom_boxplot()

cqc_teamwork <- ggplot(data, aes(x = cqc, y =  score_teamwork )) +
  geom_boxplot()

cqc_rota_design <- ggplot(data, aes(x = cqc, y =  score_rota_design )) +
  geom_boxplot()

#Descriptive analysis of survey scores
hist_survey <- ggplot(data, aes(mean_z_scores)) +
  geom_histogram(binwidth = 0.1) +
  theme_classic() +
  labs(x = "Standardised Scores",  y = "Number of Hospital Trusts", title = "Patient Surve


#scatterplot
survey_clin_sup <- ggplot(data, aes(x = score_clinical_supervision, y = mean_z_scores)) +
  geom_point() +
  theme_classic() +
```

```
  labs(x = "Junior Doctor Survey Score", y = "Standardised Patient Survey Score", title =

survey_clin_sup_ooh <- ggplot(data, aes(x = score_clinical_supervision_out_of_hours, y = m
  geom_point()

survey_teamwork <- ggplot(data, aes(x = score_teamwork, y = mean_z_scores)) +
  geom_point()

survey_rota_design <- ggplot(data, aes(x = score_rota_design, y = mean_z_scores)) +
  geom_point()


#histogram and bar charts of NTS and outcomes
nts_hist + shmi_hist + cqc_barplot + hist_survey +
  plot_layout(ncol = 2, widths = c(2, 1))
#NTs scores vs outcome variables
shmi_clin_sup + cqc_clin_sup + survey_clin_sup  +
  plot_layout(ncol = 2)
#Analysis of other predicts

#Number of trainees who responded
#hist_responses_clin_sup <- hist(data$perc_clin_sup_res)
summary_responses_clin_sup <- summary(data$perc_clin_sup_res)

#hist_responses_clin_sup_ooh <- hist(data$perc_clin_sup_ooh_res)
summary_responses_clin_sup_ooh <- summary(data$perc_clin_sup_ooh_res)

#hist_responses_team <- hist(data$perc_team_res)
summary_responses_team <- summary(data$perc_team_res)

#hist_responses_rota <- hist(data$perc_rota_res)
summary_responses_rota <- summary(data$perc_rota_res)

#res_shmi<- plot(data$perc_clin_sup_res, data$shmi_value)
#res_survey <- plot(data$perc_clin_sup_res, data$mean_z_scores)
res_cqc <- ggplot(data, aes(cqc, perc_clin_sup_res)) +
  geom_boxplot()

#ED wait time
#hist_ed <- hist(data$ed_wait)
summary_ed <- summary(data$ed_wait)
```

```r
#ed_shmi<- plot(data$ed_wait, data$shmi_value)
#ed_survey <- plot(data$ed_wait, data$mean_z_scores)
ed_cqc <- ggplot(data, aes(cqc, ed_wait)) +
  geom_boxplot()

#bed occupancy
#hist_bed <- hist(data$bed_occupancy)
summary_bed <- summary(data$bed_occupancy)
#bed_shmi<- plot(data$bed_occupancy, data$shmi_value)
#bed_survey <- plot(data$bed_occupancy, data$mean_z_scores)
bed_cqc <- ggplot(data, aes(cqc, bed_occupancy)) +
  geom_boxplot()

#financial status
##there appear to be some outliers
hist_income <- ggplot(data, aes(income)) +
  geom_histogram(bins= 1000)
summary_income <- summary(data$income)
#income_shmi<- plot(data$income, data$shmi_value)
#income_survey <- plot(data$income, data$mean_z_scores)
income_cqc <- ggplot(data, aes(cqc, income)) +
  geom_boxplot()

#number of trainees
#hist_trainees <- hist(data$total_trainees)
summary_trainees <- summary(data$total_trainees)
#trainees_shmi<- plot(data$total_trainees, data$shmi_value)
#trainees_survey <- plot(data$total_trainees, data$mean_z_scores)
trainees_cqc <- ggplot(data, aes(cqc, total_trainees)) +
  geom_boxplot()

#sickness
#hist_sickness <- hist(data$sickness_rate)
summary_sickness <- summary(data$sickness_rate)
#sickness_shmi<- plot(data$sickness_rate, data$shmi_value)
#sickness_survey <- plot(data$sickness_rate, data$mean_z_scores)
sickness_cqc <- ggplot(data, aes(cqc, sickness_rate)) +
  geom_boxplot()

#turnover
#hist_turnover <- hist(data$turnover_index)
```

```r
summary_turnover <- summary(data$turnover_index)
#turnover_shmi<- plot(data$turnover_index, data$shmi_value)
#turnover_survey <- plot(data$turnover_index, data$mean_z_scores)
turnover_cqc <- ggplot(data, aes(cqc, turnover_index)) +
  geom_boxplot()




#Check for collinearity
#pairs(data[, c("turnover_index", "total_trainees", "sickness_rate", "bed_occupancy", "ed_
#               "score_clinical_supervision_out_of_hours", "score_clinical_supervision", "
#               "score_rota_design")])

pairs(data[, c("score_clinical_supervision_out_of_hours", "score_clinical_supervision", "s
               "score_rota_design", "perc_clin_sup_res")])


#Linear modelling of SHMI
shmi_model1_clin_sup <- lm(shmi_value ~ score_clinical_supervision +
               turnover_index +
               total_trainees +
               sickness_rate +
               income +
               bed_occupancy +
                 ed_wait +
                 perc_clin_sup_res,
             data = data)
shmi_model1_clin_sup_summary <- tidy(shmi_model1_clin_sup)
shmi_model1_clin_sup_summary <- shmi_model1_clin_sup_summary %>%
  mutate(significant = ifelse(p.value < 0.05, "Y", "N"))


#flexible modelling
##quadratic
shmi_model2_clin_sup <- lm(shmi_value ~ I(score_clinical_supervision^2) +
                             score_clinical_supervision +
               turnover_index +
               total_trainees +
```

```
                sickness_rate +
                income +
                bed_occupancy +
                  ed_wait +
                  perc_clin_sup_res,
              data = data)
shmi_model2_clin_sup_summary <- tidy(shmi_model2_clin_sup)
shmi_model2_clin_sup_summary <- shmi_model2_clin_sup_summary %>%
  mutate(significant = ifelse(p.value < 0.05, "Y", "N"))

#B-spline
shmi_model3_clin_sup <- lm(shmi_value ~ bSpline(score_clinical_supervision,df=3) +
                turnover_index +
                total_trainees +
                sickness_rate +
                income +
                bed_occupancy +
                  ed_wait +
                  perc_clin_sup_res,
              data = data)
shmi_model3_clin_sup_summary <- tidy(shmi_model3_clin_sup)
shmi_model3_clin_sup_summary <- shmi_model3_clin_sup_summary %>%
  mutate(significant = ifelse(p.value < 0.05, "Y", "N"))

#Natural cubic spline
shmi_model4_clin_sup <- lm(shmi_value ~ ns(score_clinical_supervision,df=3) +
                turnover_index +
                total_trainees +
                sickness_rate +
                income +
                bed_occupancy +
                  ed_wait +
                  perc_clin_sup_res,
              data = data)
shmi_model4_clin_sup_summary <- tidy(shmi_model4_clin_sup)
shmi_model4_clin_sup_summary <- shmi_model4_clin_sup_summary %>%
  mutate(significant = ifelse(p.value < 0.05, "Y", "N"))

#GAM
shmi_model5_clin_sup <- gam(shmi_value ~ s(score_clinical_supervision,df=3) +
                turnover_index +
```

```
                total_trainees +
                sickness_rate +
                income +
                bed_occupancy +
                  ed_wait +
                  perc_clin_sup_res,
              data = data)
shmi_model5_clin_sup_summary <- tidy(shmi_model5_clin_sup)
shmi_model5_clin_sup_summary <- shmi_model5_clin_sup_summary %>%
  mutate(significant = ifelse(p.value < 0.05, "Y", "N"))

#Interaction terms
shmi_model6_clin_sup <- lm(shmi_value ~ score_clinical_supervision +
                total_trainees  +
                score_clinical_supervision:total_trainees +
                turnover_index +
                sickness_rate +
                income +
                bed_occupancy +
                  ed_wait +
                  perc_clin_sup_res,
              data = data)
shmi_model6_clin_sup_summary <- tidy(shmi_model6_clin_sup)
shmi_model6_clin_sup_summary <- shmi_model6_clin_sup_summary %>%
  mutate(significant = ifelse(p.value < 0.05, "Y", "N"))



#Reduced Model
shmi_model7_clin_sup <- lm(shmi_value ~ bSpline(score_clinical_supervision,df=3) +
                turnover_index +
                total_trainees +
                sickness_rate +
                bed_occupancy +
                  ed_wait,
              data = data)
shmi_model7_clin_sup_summary <- tidy(shmi_model7_clin_sup)
shmi_model7_clin_sup_summary <- shmi_model7_clin_sup_summary %>%
  mutate(significant = ifelse(p.value < 0.05, "Y", "N"))
```

```r
model_metrics_shmi <- data.frame(
  Model = c("Linear", "Quadratic", "B-spline", "Natural Cubic Spline", "GAM", "Reduced"),
  RMSE = c(sqrt(mean(shmi_model1_clin_sup$residuals^2)),
           sqrt(mean(shmi_model2_clin_sup$residuals^2)),
           sqrt(mean(shmi_model3_clin_sup$residuals^2)),
           sqrt(mean(shmi_model4_clin_sup$residuals^2)),
           sqrt(mean(shmi_model5_clin_sup$residuals^2)),
           sqrt(mean(shmi_model7_clin_sup$residuals^2))),
  R2 = c(summary(shmi_model1_clin_sup)$r.squared,
         summary(shmi_model2_clin_sup)$r.squared,
         summary(shmi_model3_clin_sup)$r.squared,
         summary(shmi_model4_clin_sup)$r.squared,
         NA,
         summary(shmi_model7_clin_sup)$r.squared),
  AIC = c(AIC(shmi_model1_clin_sup),
          AIC(shmi_model2_clin_sup),
          AIC(shmi_model3_clin_sup),
          AIC(shmi_model4_clin_sup),
          AIC(shmi_model5_clin_sup),
          AIC(shmi_model7_clin_sup))
)

kable(model_metrics_shmi, digits = 3)  %>%
  kable_classic()

linear_vs_spline_shmi <- anova(shmi_model1_clin_sup, shmi_model3_clin_sup)
full_vs_reduced_shmi <- anova(shmi_model3_clin_sup, shmi_model7_clin_sup)


kable(shmi_model3_clin_sup_summary, digits = 3)  %>%
  kable_classic()

#diagnostics
#plot(shmi_model3_clin_sup)
#vif(shmi_model3_clin_sup)
#Regularisation for SHMI outcome data
X <- model.matrix( ~ score_clinical_supervision +
                     score_clinical_supervision_out_of_hours +
                     score_teamwork +
                     score_rota_design +
                turnover_index +
```

```
                total_trainees +
                sickness_rate +
                income +
                bed_occupancy +
                  ed_wait +
                  perc_clin_sup_res,
              data = data)[,-1]
y <- data$shmi_value[complete.cases(data$ed_wait)]

ridge_model <- glmnet(X, y, alpha = 0, family = "gaussian")
lasso_model <- glmnet(X, y, alpha = 1, family = "gaussian")
elasticnet_model <- glmnet(X, y, alpha = 0.5, family = "gaussian")

ridge_model_cv <- cv.glmnet(X, y, alpha = 0, family = "gaussian")
lasso_model_cv <- cv.glmnet(X, y, alpha = 1, family = "gaussian")
elasticnet_model_cv <- cv.glmnet(X, y, alpha = 0.5, family = "gaussian")


par(mfrow=c(1, 3))
plot(ridge_model, xvar = "lambda", label = TRUE)
plot(lasso_model, xvar = "lambda", label = TRUE)
plot(elasticnet_model, xvar = "lambda", label = TRUE)

plot(ridge_model_cv, xvar = "lambda", label = TRUE)
plot(lasso_model_cv, xvar = "lambda", label = TRUE)
plot(elasticnet_model_cv, xvar = "lambda", label = TRUE)

optimal_lambda_ridge <- ridge_model_cv$lambda.min
ridge_coefs <- as.matrix(coef(ridge_model, s = optimal_lambda_ridge))
optimal_lambda_lasso <- lasso_model_cv$lambda.min
lasso_coefs <- as.matrix(coef(lasso_model_cv, s = optimal_lambda_lasso))
optimal_lambda_elastic <- elasticnet_model_cv$lambda.min
elastic_coefs <- as.matrix(coef(elasticnet_model, s = optimal_lambda_elastic))


summary <- data.frame(ridge_coefs, lasso_coefs, elastic_coefs)
colnames(summary) = c("ridge", "lasso", "elastic")
kable(summary, digits = 3) %>%
  kable_classic()
```

```r
#Linear modelling of IP satisfaction survey
survey_model1_clin_sup <- lm(mean_z_scores ~ score_clinical_supervision +
               turnover_index +
               total_trainees +
               sickness_rate +
               income +
               bed_occupancy +
                 ed_wait +
                 perc_clin_sup_res,
             data = data)
survey_model1_clin_sup_summary <- tidy(survey_model1_clin_sup)
survey_model1_clin_sup_summary <- survey_model1_clin_sup_summary %>%
  mutate(significant = ifelse(p.value < 0.05, "Y", "N"))

#Flexible modelling
##Quadratic
survey_model2_clin_sup <- lm(mean_z_scores ~ score_clinical_supervision +
                I(score_clinical_supervision^2) +
               turnover_index +
               total_trainees +
               sickness_rate +
               income +
               bed_occupancy +
                 ed_wait +
                 perc_clin_sup_res,
             data = data)
survey_model2_clin_sup_summary <- tidy(survey_model2_clin_sup)
survey_model2_clin_sup_summary <- survey_model2_clin_sup_summary %>%
  mutate(significant = ifelse(p.value < 0.05, "Y", "N"))

#B-spline
survey_model3_clin_sup <- lm(mean_z_scores ~ bSpline(score_clinical_supervision, df =3) +
               turnover_index +
               total_trainees +
               sickness_rate +
               income +
               bed_occupancy +
                 ed_wait +
                 perc_clin_sup_res,
             data = data)
survey_model3_clin_sup_summary <- tidy(survey_model3_clin_sup)
```

```r
survey_model3_clin_sup_summary <- survey_model3_clin_sup_summary %>%
  mutate(significant = ifelse(p.value < 0.05, "Y", "N"))

#natural cubic spline
survey_model4_clin_sup <- lm(mean_z_scores ~ ns(score_clinical_supervision, df =3) +
               turnover_index +
               total_trainees +
               sickness_rate +
               income +
               bed_occupancy +
                 ed_wait +
                 perc_clin_sup_res,
             data = data)
survey_model4_clin_sup_summary <- tidy(survey_model4_clin_sup)
survey_model4_clin_sup_summary <- survey_model3_clin_sup_summary %>%
  mutate(significant = ifelse(p.value < 0.05, "Y", "N"))

#GAM
survey_model5_clin_sup <- gam(mean_z_scores ~ s(score_clinical_supervision, df =3) +
               turnover_index +
               total_trainees +
               sickness_rate +
               income +
               bed_occupancy +
                 ed_wait +
                 perc_clin_sup_res,
             data = data)
survey_model5_clin_sup_summary <- tidy(survey_model5_clin_sup)
survey_model5_clin_sup_summary <- survey_model5_clin_sup_summary %>%
  mutate(significant = ifelse(p.value < 0.05, "Y", "N"))

#Interaction terms
survey_model6_clin_sup <- lm(mean_z_scores ~ ns(score_clinical_supervision, df = 3) +
               total_trainees  +
               score_clinical_supervision:turnover_index +
               turnover_index +
               sickness_rate +
               income +
               bed_occupancy +
                 ed_wait +
                 perc_clin_sup_res,
```

```
              data = data)
survey_model6_clin_sup_summary <- tidy(survey_model6_clin_sup)
survey_model6_clin_sup_summary <- survey_model6_clin_sup_summary %>%
  mutate(significant = ifelse(p.value < 0.05, "Y", "N"))



#Reduced Model
survey_model7_clin_sup <- lm(mean_z_scores ~ ns(score_clinical_supervision,df=3) +
                turnover_index +
                total_trainees +
                sickness_rate +
                bed_occupancy +
                  ed_wait,
              data = data)
survey_modely_clin_sup_summary <- tidy(survey_model7_clin_sup)
survey_modely_clin_sup_summary <- survey_modely_clin_sup_summary %>%
  mutate(significant = ifelse(p.value < 0.05, "Y", "N"))


#model metrics
model_metrics_survey <- data.frame(
  Model = c("Linear", "Quadratic", "B-spline", "Natural Cubic Spline", "GAM", "Reduced", "
  RMSE = c(sqrt(mean(survey_model1_clin_sup$residuals^2)),
           sqrt(mean(survey_model2_clin_sup$residuals^2)),
           sqrt(mean(survey_model3_clin_sup$residuals^2)),
           sqrt(mean(survey_model4_clin_sup$residuals^2)),
           sqrt(mean(survey_model5_clin_sup$residuals^2)),
           sqrt(mean(survey_model7_clin_sup$residuals^2)),
           sqrt(mean(survey_model6_clin_sup$residuals^2))),
  R2 = c(summary(survey_model1_clin_sup)$r.squared,
         summary(survey_model2_clin_sup)$r.squared,
         summary(survey_model3_clin_sup)$r.squared,
         summary(survey_model4_clin_sup)$r.squared,
         NA,
         summary(survey_model7_clin_sup)$r.squared,
         summary(survey_model6_clin_sup)$r.squared),
  AIC = c(AIC(survey_model1_clin_sup),
          AIC(survey_model2_clin_sup),
          AIC(survey_model3_clin_sup),
          AIC(survey_model4_clin_sup),
```

```
          AIC(survey_model5_clin_sup),
          AIC(survey_model7_clin_sup),
          AIC(survey_model6_clin_sup))
)



kable(model_metrics_survey, digits = 3) %>%
  kable_classic()

linear_vs_spline_survey <- anova(survey_model1_clin_sup, survey_model4_clin_sup)
full_vs_reduced_survey <- anova(survey_model6_clin_sup, survey_model7_clin_sup)
kable(survey_model6_clin_sup_summary, digits = 3) %>%
  kable_classic()

#diagnostics
#plot(survey_model6_clin_sup)
#vif(survey_model4_clin_sup)
#Regularisation for IP survey outcome data
X <- model.matrix( ~ score_clinical_supervision +
                     score_clinical_supervision:turnover_index +
                     score_clinical_supervision_out_of_hours +
                     score_teamwork +
                     score_rota_design +
                turnover_index +
                total_trainees +
                sickness_rate +
                income +
                bed_occupancy +
                  ed_wait +
                  perc_clin_sup_res,
              data = data)[,-1]
y <- data$mean_z_scores[complete.cases(data$ed_wait)]

ridge_model <- glmnet(X, y, alpha = 0, family = "gaussian")
lasso_model <- glmnet(X, y, alpha = 1, family = "gaussian")
elasticnet_model <- glmnet(X, y, alpha = 0.5, family = "gaussian")

ridge_model_cv <- cv.glmnet(X, y, alpha = 0, family = "gaussian")
lasso_model_cv <- cv.glmnet(X, y, alpha = 1, family = "gaussian")
elasticnet_model_cv <- cv.glmnet(X, y, alpha = 0.5, family = "gaussian")
```

```r
par(mfrow=c(1, 3))
plot(ridge_model, xvar = "lambda", label = TRUE)
plot(lasso_model, xvar = "lambda", label = TRUE)
plot(elasticnet_model, xvar = "lambda", label = TRUE)

plot(ridge_model_cv, xvar = "lambda", label = TRUE)
plot(lasso_model_cv, xvar = "lambda", label = TRUE)
plot(elasticnet_model_cv, xvar = "lambda", label = TRUE)

optimal_lambda_ridge <- ridge_model_cv$lambda.min
ridge_coefs <- as.matrix(coef(ridge_model, s = optimal_lambda_ridge))
optimal_lambda_lasso <- lasso_model_cv$lambda.min
lasso_coefs <- as.matrix(coef(lasso_model_cv, s = optimal_lambda_lasso))
optimal_lambda_elastic <- elasticnet_model_cv$lambda.min
elastic_coefs <- as.matrix(coef(elasticnet_model, s = optimal_lambda_elastic))

summary <- data.frame(ridge_coefs, lasso_coefs, elastic_coefs)
colnames(summary) = c("ridge", "lasso", "elastic")
kable(summary, digits = 3) %>%
  kable_classic()


#Ordinal regression modelling of CQC rating
cqc_model1_clin_sup <- vglm(cqc ~ score_clinical_supervision +
               turnover_index +
               total_trainees +
               sickness_rate +
               income +
               bed_occupancy +
               ed_wait +
               perc_clin_sup_res,
               cumulative(parallel=TRUE, reverse=TRUE),
                   data=data)

summary(cqc_model1_clin_sup)



#reduced
cqc_model2_clin_sup <- vglm(cqc ~ score_clinical_supervision +
```

```
                sickness_rate +
                income +
                bed_occupancy +
                ed_wait +
                perc_clin_sup_res,
                cumulative(parallel=TRUE, reverse=TRUE),
                    data=data)


##could change to binary
#data$cqc_binary <- ifelse(data$cqc %in% c("Outstanding", "Good"), 1, 0)
```