Group 13

Liying Choo, A0172633H; Lim Kay Yee Rachel, A0114716L

2025-11-20

```
packages <- c("sf",
              "sp",
              "spatstat",
              "gstat",
              "spdep",
              "nlme",
              "SpatialEpi",
              "spatialreg",
              "terra",
              "ggplot2",
              "dplyr",
              "purrr",
              "geodata",
              "tidyterra",
              "patchwork",
              "knitr",
              "tidyverse",
              "httr",
              "broom",
              "kableExtra",
              "car",
              "webshot",
              "webshot2")

lapply(packages, library, character.only = TRUE)
rm(packages)
```

Part I: Analysis of Lip Cancer Data in Scotland

Read data

```
scotland_sf <- readRDS("Scotland.rds")
```

1(a)

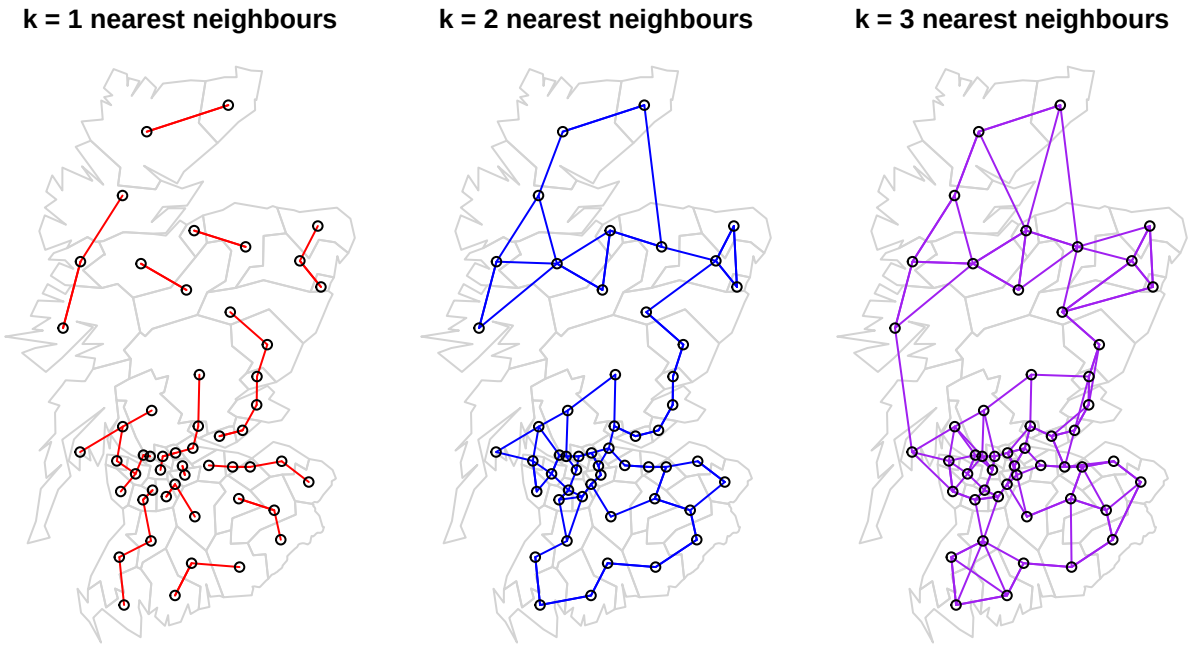
```r
center <- st_centroid(scotland_sf)
```

```
## Warning: st_centroid assumes attributes are constant over geometries
```

```r
# set the number of nearest neighbors to be 1, 2 and 3
knn_1 <- knn2nb(knearneigh(center, k = 1))
```

```
## Warning in knn2nb(knearneigh(center, k = 1)): neighbour object has 14
## sub-graphs
```

```r
knn_2 <- knn2nb(knearneigh(center, k = 2))
knn_3 <- knn2nb(knearneigh(center, k = 3))

par(mfrow=c(1,3), mai = c(1, 0.1, 0.1, 0.1))

plot(st_geometry(scotland_sf), border = "lightgray")
plot.nb(knn_1, st_geometry(scotland_sf), col= "red", add = TRUE)
title("k = 1 nearest neighbours")

plot(st_geometry(scotland_sf), border = "lightgray")
plot.nb(knn_2, st_geometry(scotland_sf), col= "blue", add = TRUE)
title("k = 2 nearest neighbours")

plot(st_geometry(scotland_sf), border = "lightgray")
plot.nb(knn_3, st_geometry(scotland_sf), col= "purple", add = TRUE)
title("k = 3 nearest neighbours")
```




```
# Check disconnected subgraphs
n.comp.nb(knn_1)$nc
```

```
## [1] 14
```

```
n.comp.nb(knn_2)$nc
```

```
## [1] 1
```

```
n.comp.nb(knn_3)$nc
```

```
## [1] 1
```

From a visual comparison, as well as the check on disconnected subgraphs, the regions are all connected when the k-nearest neighbours are 2 or 3, but are not connected in the plot for 1 nearest neighbour.

1(b)

```
nb_queen <- poly2nb(scotland_sf, queen = TRUE)
head(nb_queen)
```

```
## [[1]]
## [1]  5  7 16
##
## [[2]]
## [1] 6 8
##
## [[3]]
## [1] 9
##
## [[4]]
## [1] 15 17 25
##
## [[5]]
## [1]  1  9 16
##
## [[6]]
## [1]  2  8 10 13 14
```

```
nb_rook <- poly2nb(scotland_sf, queen = FALSE)
head(nb_rook)
```

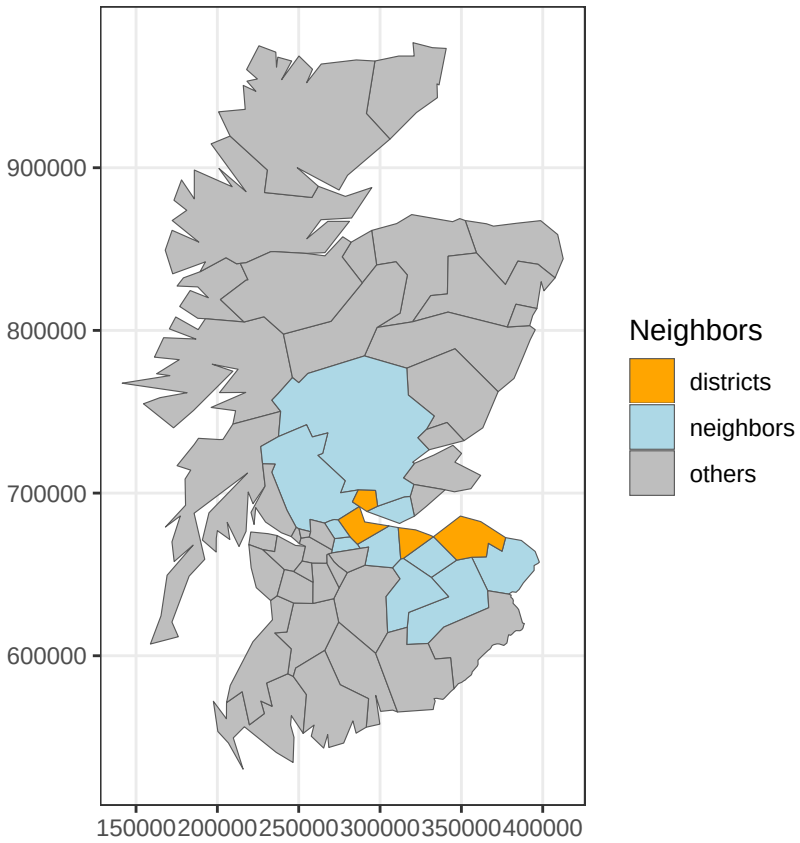
```
## [[1]]
## [1]  5  7 16
##
## [[2]]
## [1] 6 8
##
## [[3]]
## [1] 9
##
## [[4]]
## [1] 15 17 25
##
## [[5]]
## [1]  1  9 16
##
## [[6]]
## [1]  2  8 10 13 14
```

```
### detecting regions with differences
nb_diff <- mapply("%in%", nb_queen, nb_rook)
nb_diff_id <- nb_diff %>%
  map_lgl(~ any(!.x)) %>%
  which()

scotland_sf$name[nb_diff_id]
```

```
## [1] EastLothian Falkirk     Clackmannan Edinburgh
## 56 Levels: Aberdeen Angus Annandale Argyll-Bute Badenoch ... Wigtown
```

The 4 districts are EastLothian, Falkirk, Clackmannan and Edinburgh.

```r
# add variable neigh_diff
scotland_sf <- scotland_sf %>%
  mutate(neigh_diff = "others")

# add districts
scotland_sf$neigh_diff[nb_diff_id] <- "districts"

# add neighbours
neighbours <- unique(unlist(nb_queen[nb_diff_id]))
neighbours <- setdiff(neighbours, nb_diff_id)
scotland_sf$neigh_diff[neighbours] <- "neighbors"

#plot
ggplot(scotland_sf) +
  geom_sf(aes(fill = neigh_diff)) +
 scale_fill_manual(
   name   = "Neighbors",
   breaks = c("districts", "neighbors", "others"),
   values = c("orange", "lightblue", "grey")) +
  theme_bw()
```



1(c)

```
b_weight <- nb2listw(nb_rook, style="B")

# normal test (without skewness correction)
g_moran <- moran.test(scotland_sf$cancer, b_weight, randomisation = FALSE, alternative = "greater")

# permutation test
set.seed(5226)
g_moran_mc <- moran.mc(scotland_sf$cancer, b_weight, nsim = 9999)

# Monte Carlo test
moran.pois <- function(y, n_vec, listw, nsim) {
 Tstat <- rep(0, nsim)
 Tstat[1] <- moran(y, listw, length(y), Szero(listw))$I
 cr <- sum(y)/sum(n_vec)
 pmeans <- cr*n_vec

 for(ii in 2:nsim) {
 tmp <- rpois(rep(1, length(pmeans)), pmeans)
 Tstat[ii] <- moran(tmp, listw, length(y), Szero(listw))$I
 }
 formatC(
  sum(Tstat[-1] > Tstat[1])/(nsim+1),
  format = "f",
  digits = 3)
}

set.seed(5226)
g_moran_poisson <- moran.pois(scotland_sf$cancer, scotland_sf$expected, b_weight, 9999)
```

```
g_moran
```

```
##
##  Moran I test under normality
##
## data:  scotland_sf$cancer
## weights: b_weight
##
## Moran I statistic standard deviate = 1.8105, p-value = 0.03511
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##       0.139454218      -0.019230769       0.007681815
```

```
g_moran_mc
```

```
##
##  Monte-Carlo simulation of Moran I
##
## data:  scotland_sf$cancer
## weights: b_weight
```

```
## number of simulations + 1: 10000
##
## statistic = 0.13945, observed rank = 9556, p-value = 0.0444
## alternative hypothesis: greater
```

```
g_moran_poisson
```

```
## [1] "0.000"
```

All tests above shared outputs of p-values that are smaller than .05. This shows that there exists positive spatial dependence for lip cancer.

1(d)

```
# normal test
geary_normal <- geary.test(scotland_sf$cancer, b_weight, randomisation = FALSE)

# permutation test
geary_mc <- geary.mc(scotland_sf$cancer, b_weight, nsim = 9999)

# Monte Carlo test
geary.pois <- function(y, n_vec, listw, nsim) {
 Tstat <- rep(0, nsim)
 Tstat[1] <- geary(y, listw, length(y), length(y)-1,
 Szero(listw))$C
 cr <- sum(y)/sum(n_vec)
 pmeans <- cr*n_vec

 for(ii in 2:nsim) {
 tmp <- rpois(rep(1, length(pmeans)), pmeans)
 Tstat[ii] <- geary(tmp, listw, length(y), length(y)-1,
 Szero(listw))$C
 }
 formatC(
  sum(Tstat[-1] < Tstat[1])/(nsim+1),
  format = "f",
  digits = 3)
 }

set.seed(5226)
geary_poisson <- geary.pois(scotland_sf$cancer, scotland_sf$expected, b_weight, 9999)
```

```
geary_normal
```

```
##
##  Geary C test under normality
##
## data:  scotland_sf$cancer
## weights: b_weight
##
```

```
## Geary C statistic standard deviate = 2.0026, p-value = 0.02261
## alternative hypothesis: Expectation greater than statistic
## sample estimates:
## Geary C statistic       Expectation          Variance
##        0.75393886        1.00000000        0.01509767
```

```
geary_mc
```

```
##
##  Monte-Carlo simulation of Geary C
##
## data:  scotland_sf$cancer
## weights: b_weight
## number of simulations + 1: 10000
##
## statistic = 0.75394, observed rank = 416, p-value = 0.0416
## alternative hypothesis: greater
```

```
geary_poisson
```

```
## [1] "0.000"
```

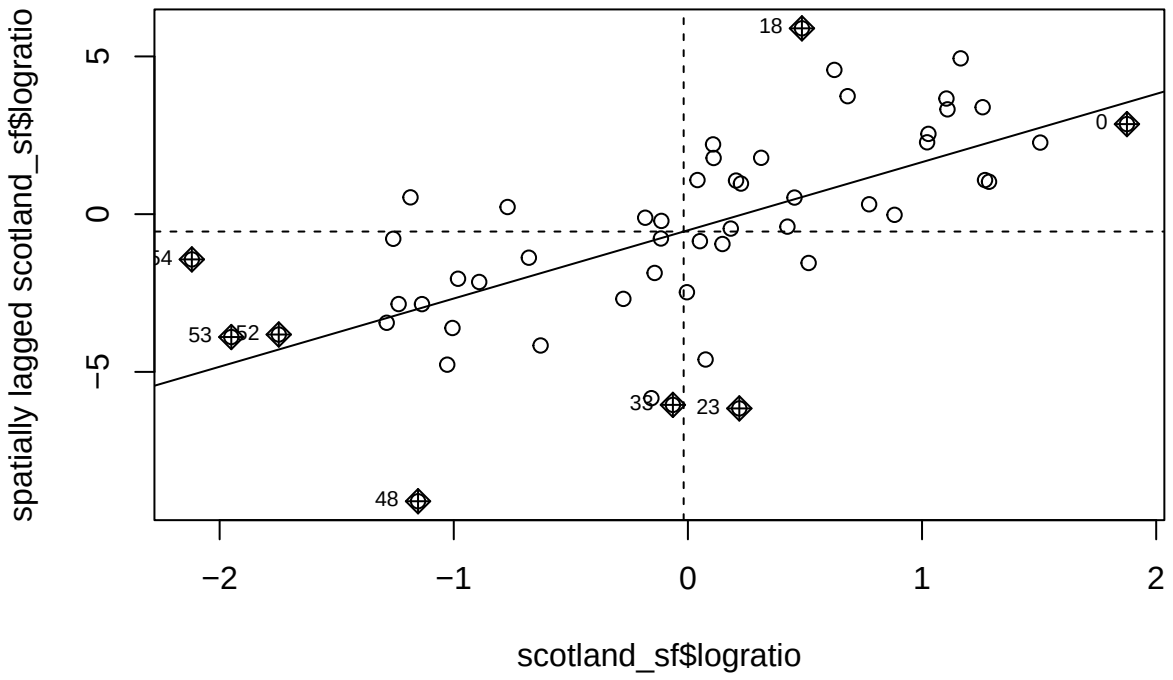
The results from the equivalent Geary's tests are similar to the tests for Moran's I, with p-values that are smaller than .05. The conclusion remains the same, that there exists positive spatial dependence for lip cancer.

1(e)

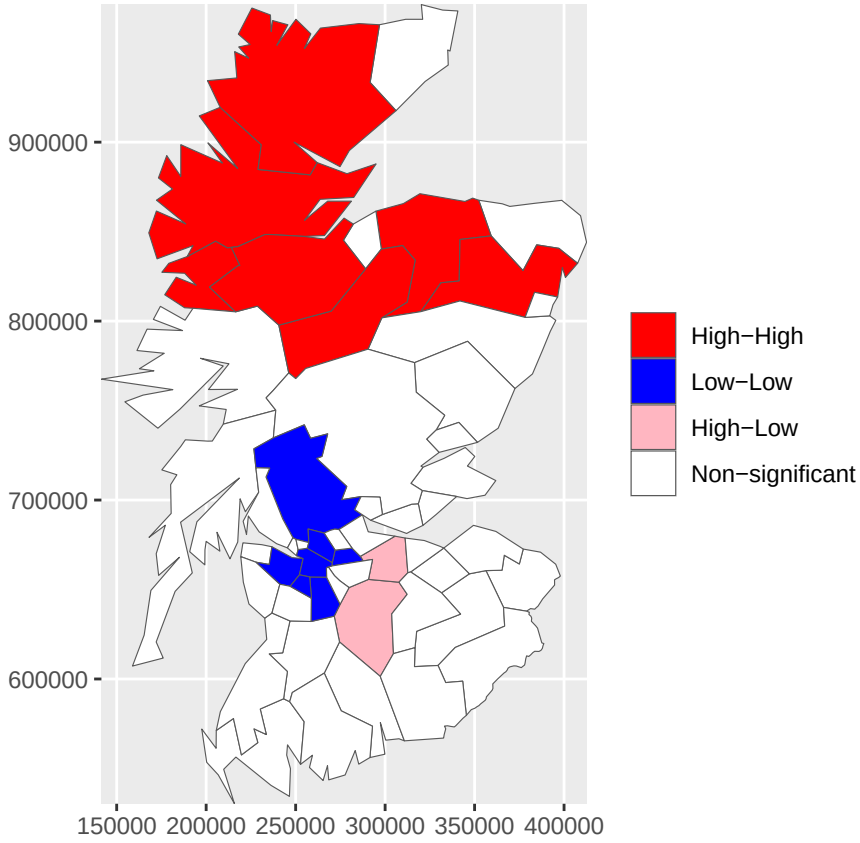
```
loc_moran <- localmoran(scotland_sf$logratio, b_weight, alternative = "two.sided")
head(loc_moran)
```

```
##         Ii        E.Ii    Var.Ii      Z.Ii Pr(z != E(Ii))
## 0 6.406050 -0.24054605 11.266834 1.9801512     0.04768654
## 1 4.084043 -0.10376890  5.112157 1.8521879     0.06399885
## 2 1.585801 -0.03804668  1.939754 1.1659288     0.24364321
## 3 1.699379 -0.11117816  5.451559 0.7754465     0.43807585
## 4 5.115054 -0.10941890  5.368562 2.2548281     0.02414412
## 6 6.920763 -0.15656092  7.407485 2.6003611     0.00931257
```

```
mp <- moran.plot(scotland_sf$logratio, b_weight)
```




```
scotland_sf$quadrant <- NA
p.2side <- loc_moran[, "Pr(z != E(Ii))"]

# high-high
scotland_sf[(mp$x >= 0 & mp$wx >= 0) & (p.2side <= 0.05), "quadrant"] <- 1

# low-low
scotland_sf[(mp$x <= 0 & mp$wx <= 0) & (p.2side <= 0.05), "quadrant"] <- 2

# high-low
scotland_sf[(mp$x >= 0 & mp$wx <= 0) & (p.2side <= 0.05), "quadrant"] <- 3

# low-high
scotland_sf[(mp$x <= 0 & mp$wx >= 0) & (p.2side <= 0.05), "quadrant"] <- 4

# non-significant
scotland_sf[(p.2side > 0.05), "quadrant"] <- 5

ggplot(data=scotland_sf) +
  geom_sf(aes(fill=as.factor(quadrant))) +
 coord_sf(expand = FALSE) +
 scale_fill_manual(breaks=c(1,2,3,4,5),
 labels = c("High-High", "Low-Low", "High-Low", "Low-High", "Non-significant"),
 values=c("red", "blue", "lightpink", "skyblue2", "white")) +
 theme(legend.title=element_blank())
```



2(a)

```
model_1 <- gls(logratio ~ percentAFF + northkm + eastkm,
data = scotland_sf, method="ML")

model_1 %>% summary()
```

```
## Generalized least squares fit by maximum likelihood
##   Model: logratio ~ percentAFF + northkm + eastkm
##   Data: scotland_sf
##        AIC     BIC    logLik
##   116.4535 126.305 -53.22677
##
## Coefficients:
##                  Value  Std.Error   t-value p-value
## (Intercept) -0.5063387 0.15972718 -3.170022  0.0026
## percentAFF   0.0582723 0.01471083  3.961184  0.0002
## northkm      0.0051925 0.00107286  4.839900  0.0000
## eastkm      -0.0001017 0.00198195 -0.051309  0.9593
##
##  Correlation:
##             (Intr) prcAFF nrthkm
## percentAFF -0.706
```

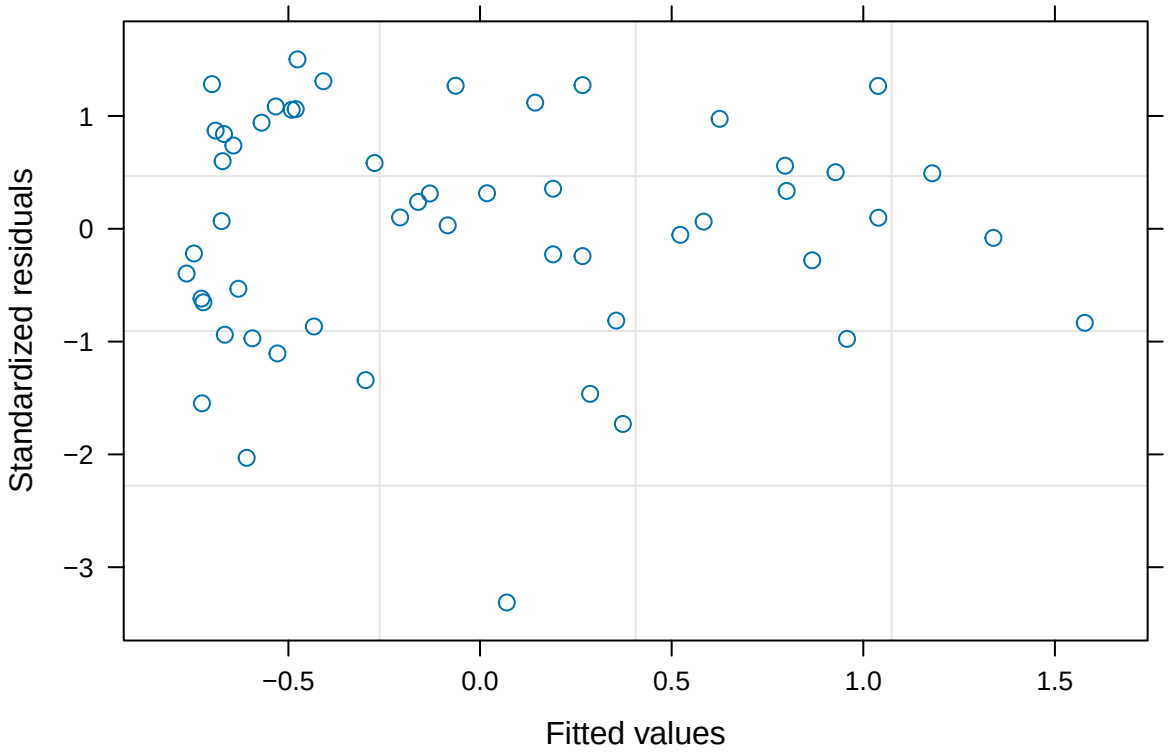
```
## northkm     0.150 -0.148
## eastkm     -0.258 -0.178 -0.088
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -3.30925516 -0.65419151  0.09174238  0.83200101  1.51117150
##
## Residual standard error: 0.6605649
## Degrees of freedom: 53 total; 49 residual
```

From the summary of the initial model fit, we can see that percentAFF and northkm are significant predictors (p-value < 0.05), while eastkm is not significant (p-value = 0.9593 > 0.05). Therefore, we will drop eastkm from the model and refit the model with only percentAFF and northkm as predictors.

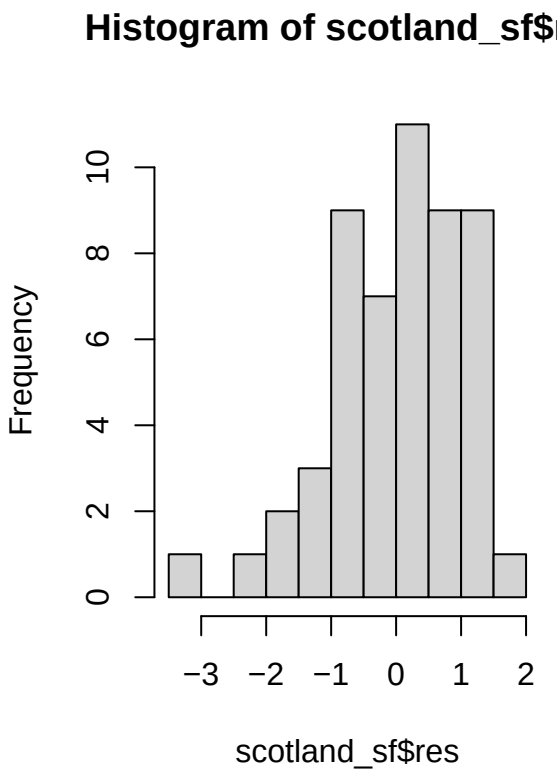
```
model_2 <- gls(logratio ~ percentAFF + northkm,
data = scotland_sf, method="ML")
model_2 %>% summary()
```

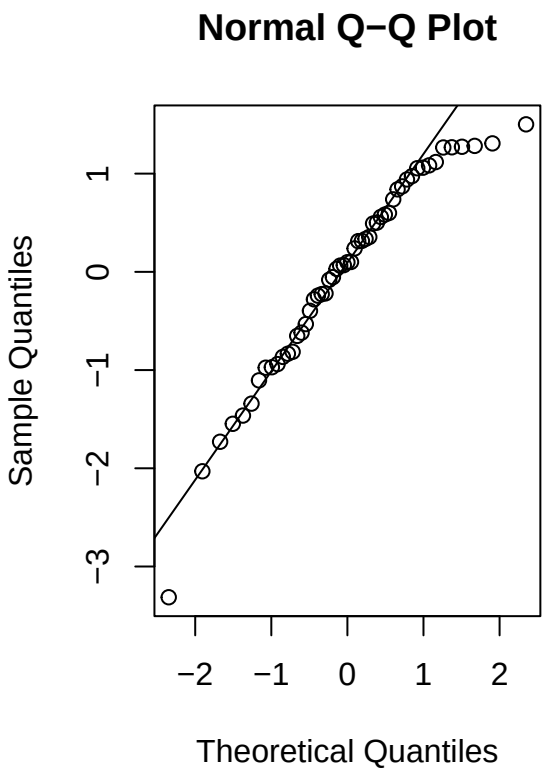
```
## Generalized least squares fit by maximum likelihood
##   Model: logratio ~ percentAFF + northkm
##   Data: scotland_sf
##        AIC      BIC    logLik
##   114.4564 122.3376 -53.22819
##
## Coefficients:
##                  Value  Std.Error   t-value p-value
## (Intercept) -0.5084557 0.15275911 -3.328481  0.0016
## percentAFF   0.0581379 0.01433056  4.056918  0.0002
## northkm      0.0051877 0.00105794  4.903558  0.0000
##
##  Correlation:
##            (Intr) prcAFF
## percentAFF -0.791
## northkm     0.132 -0.167
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -3.31295026 -0.65198697  0.09813454  0.83987636  1.50175251
##
## Residual standard error: 0.6605827
## Degrees of freedom: 53 total; 50 residual
```

```
plot(model_2)
```




```
scotland_sf$res <- residuals(model_2, type ="pearson")
opar <- par(mfrow=c(1,2))
hist(scotland_sf$res)
qqnorm(scotland_sf$res)
qqline(scotland_sf$res)
```






```
par(opar)
```

From the 'Residuals vs Fitted' and 'Scale-Location' plots, we can see that the residuals are randomly scattered around zero, indicating that there is no significant violation of the homogeneous variance assumption. The 'Q-Q Residuals' plot shows that the residuals lie close to the 45 degree reference line; and as such, are approximately normally distributed, indicating that there is no significant violation of the normality assumption. Overall, the model seems to satisfy both assumptions.

2(b)

```
wls <- gls(logratio ~ percentAFF + northkm,
data = scotland_sf,
method="ML",
weights = varFixed(~(1/expected)))
wls %>% summary()
```

```
## Generalized least squares fit by maximum likelihood
##   Model: logratio ~ percentAFF + northkm
##   Data: scotland_sf
##        AIC      BIC    logLik
##   119.2159 127.0971 -55.60797
##
## Variance function:
```

```
##  Structure: fixed weights
##  Formula: ~(1/expected)
##
## Coefficients:
##                  Value  Std.Error   t-value p-value
## (Intercept) -0.6600671 0.11648012 -5.666779   0e+00
## percentAFF   0.0617328 0.01380952  4.470308   0e+00
## northkm      0.0047140 0.00130326  3.617090   7e-04
##
##  Correlation:
##             (Intr) prcAFF
## percentAFF -0.719
## northkm     0.387 -0.364
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -2.4481191 -0.5009908  0.3129100  0.8108233  1.8674490
##
## Residual standard error: 1.77167
## Degrees of freedom: 53 total; 50 residual
```

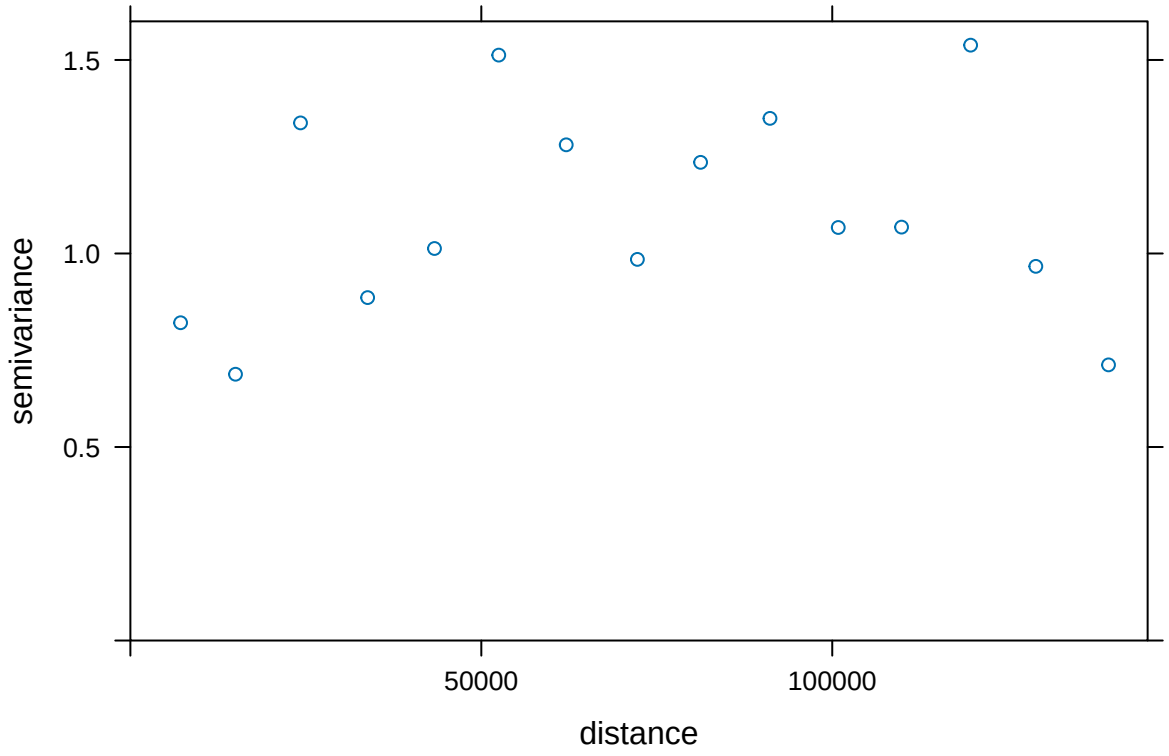
```
anova(model_1, model_2, wls) %>%
arrange(AIC)
```

```
##         Model df      AIC      BIC    logLik   Test     L.Ratio p-value
## model_2     2  4 114.4564 122.3375 -53.22819 1 vs 2 0.002847467  0.9574
## model_1     1  5 116.4535 126.3050 -53.22677
## wls         3  4 119.2159 127.0971 -55.60797
```

Both the OLS and WLS models seem to satisfy the assumptions of homogeneous variance and normality based on the diagnostic plots. However, the OLS model (model_2) has smaller AIC and BIC values compared to the WLS model, indicating that the OLS model provides an overall better model fit to the data. Therefore, we will choose the OLS model (model_2) as the better model.

2(c)

```
vgm <- variogram(res ~ 1, scotland_sf)
plot(vgm)
```




```
gls <- gls(logratio ~ percentAFF + northkm,

data = scotland_sf,
method = "ML",
corr = corExp(value = c(5000, 0.5),
form = ~easting+northing,
nugget = TRUE))

gls %>% summary()
```

```
## Generalized least squares fit by maximum likelihood
##   Model: logratio ~ percentAFF + northkm
##   Data: scotland_sf
##        AIC      BIC    logLik
##   116.2772 128.0989 -52.13858
##
## Correlation Structure: Exponential spatial correlation
##  Formula: ~easting + northing
##  Parameter estimate(s):
##        range       nugget
## 1.006668e+04 9.550765e-02
##
## Coefficients:
##                    Value  Std.Error   t-value p-value
## (Intercept) -0.29337234 0.18813511 -1.559371  0.1252
```

```
## percentAFF   0.04432762 0.01559905  2.841686  0.0065
## northkm      0.00488425 0.00110187  4.432696  0.0001
##
##  Correlation:
##            (Intr) prcAFF
## percentAFF -0.811
## northkm    -0.008 -0.082
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -3.34199640 -0.85419403 -0.00769754  0.56891769  1.33248800
##
## Residual standard error: 0.6592406
## Degrees of freedom: 53 total; 50 residual
```

Among the three models (OLS, WLS, and GLS), the OLS model without weighted variance and without spatially dependent residuals remains the best fit to the data, as it has the lowest AIC and BIC values amongst the three models. Furthermore, spatial correlation through the exponential semivariogram indicated an increase in both AIC and BIC relative to the OLS model (model_2), indicating that model fit worsened and the residuals are not strongly spatially dependent.

2(d)

```
nb_rook <- poly2nb(as_Spatial(st_geometry(scotland_sf)), queen = FALSE)
lwB <- nb2listw(nb_rook, style = "B", zero.policy = TRUE)

sar <- spautolm(logratio ~ percentAFF + northkm,
data = scotland_sf, listw = lwB)
sar %>% summary()
```

```
##
## Call: spautolm(formula = logratio ~ percentAFF + northkm, data = scotland_sf,
##     listw = lwB)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.16110 -0.47595  0.12558  0.48523  0.99715
##
## Coefficients:
##               Estimate Std. Error z value  Pr(>|z|)
## (Intercept) -0.4092391  0.1680313 -2.4355 0.0148715
## percentAFF   0.0478325  0.0144586  3.3082 0.0009389
## northkm      0.0052069  0.0011804  4.4113 1.027e-05
##
## Lambda: 0.046519 LR test value: 0.61577 p-value: 0.43262
## Numerical Hessian standard error of lambda: 0.055727
##
## Log likelihood: -52.92031
## ML residual variance (sigma squared): 0.42703, (sigma: 0.65348)
## Number of observations: 53
## Number of parameters estimated: 5
## AIC: 115.84
```

```r
set.seed(5226)
perm_1 <- moran.mc(residuals(sar), lwB, nsim = 9999)
perm_1
```

```
##
##  Monte-Carlo simulation of Moran I
##
## data:  residuals(sar)
## weights: lwB
## number of simulations + 1: 10000
##
## statistic = -0.0029902, observed rank = 5876, p-value = 0.4124
## alternative hypothesis: greater
```

The summary of the SAR model shows that the likelihood ratio test of the spatial dependence parameter (ρ) is not statistically significant (ρ = 0.433). The permutation test using Moran's I also shows a large p-value of 0.412. As such the two conclusions agree with one another, and we can conclude that there is no significant spatial dependence amongst the residuals.

2(e)

```r
car <- spautolm(logratio ~ percentAFF + northkm,
data = scotland_sf, listw = lwB, family = "CAR")
car %>% summary()
```

```
##
## Call: spautolm(formula = logratio ~ percentAFF + northkm, data = scotland_sf,
##     listw = lwB, family = "CAR")
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.19864 -0.46406  0.12476  0.44575  1.01998
##
## Coefficients:
##               Estimate Std. Error z value  Pr(>|z|)
## (Intercept) -0.3993282  0.1707611 -2.3385  0.019360
## percentAFF   0.0473795  0.0144754  3.2731  0.001064
## northkm      0.0051664  0.0011896  4.3429 1.406e-05
##
## Lambda: 0.090028 LR test value: 0.63319 p-value: 0.42619
## Numerical Hessian standard error of lambda: 0.096531
##
## Log likelihood: -52.9116
## ML residual variance (sigma squared): 0.42241, (sigma: 0.64993)
## Number of observations: 53
## Number of parameters estimated: 5
## AIC: 115.82
```

```r
set.seed(5226)
perm_2 <- moran.mc(residuals(car), lwB, nsim = 9999)
perm_2
```

```
##
##  Monte-Carlo simulation of Moran I
##
## data:  residuals(car)
## weights: lwB
## number of simulations + 1: 10000
##
## statistic = -0.098752, observed rank = 1748, p-value = 0.8252
## alternative hypothesis: greater
```

No, the results of the one-parameter CAR model does not change the conclusion made previously about the spatial dependence in the residuals. This is because the permutation test using Moran's I (I = -0.098752, p = 0.825) show the same conclusions - that there is a negative spatial dependence in the residuals, but that the spatial lag parameter (ρ) was not statistically significant as $p > 0.05$. As such, it supports the previous conclusion made that there is no significant evidence of (positive) spatial dependence in the residuals.

PART II: Course Project

Introduction

Data Overview

1. resale (Data Type: Attribute; Source: https://data.gov.sg/datasets?topics=housing&query=hdb+price&resultId=d_8b84c4ee58e3cfc0ece0d773c8ca6abc)
2. sg_boundary (Data Type: Attribute; Source: https://data.gov.sg/datasets?query=subzone&resultId=d_8594ae9ff96d0c708bc2af633048edfb)
3. hdb (Data Type: Attribute; Source: https://github.com/BlueSkyLT/siteselect_sg)
4. CHASClinics (Data Type: Attribute; Source: https://data.gov.sg/datasets/d_548c33ea2d99e29ec63a7cc9edcccedc/view)
5. LTAMRTStationExitGEOJSON (Data Type: Attribute; Source: https://data.gov.sg/datasets?query=lta+mrt+station+exit&resultId=d_b39d3a0871985372d7e1637193335da5&sidebar=false)
6. LTASchoolZone (Data Type: Attribute; Source: https://data.gov.sg/datasets?query=lta+school+zoner&resultId=d_abf023b38d9bc451484e3d67b562bc5c&sidebar=false)
7. NEAMarketandFoodCentre (Data Type: Attribute; Source: https://data.gov.sg/datasets?query=nea+market+and+food+cent&resultId=d_a57a245b3cf3ec76ad36d55393a16e97)
8. Parks (Data Type: Attribute; Source: https://data.gov.sg/datasets/d_0542d48f0991541706b58059381a6eca/view)

Problem to be investigated and statistical methods used

We are analysing spatial patterns in HDB resale prices from 2017-2025 across Singapore.

Research Question:

1. Is the strength of the relationship between HDB flat size and resale prices spatially clustered across Singapore's planning areas?

   - **Statistical_Method**: Global Moran's I on area-specific size–price regression coefficients

2. Which planning areas show stronger or weaker links between flat size and resale price, compared to their neighbouring planning areas?

   - **Statistical_Method**: Local Moran's I for on area-specific size–price regression coefficients

3. Which specific amenity features (amongst MRT, school, retail, park, and healthcare proximity) are the strongest factors in explaining variation in the marginal willingness to pay for additional floor area (beta_size_price) across Singapore's planning areas?

   - **Statistical_Method**: OLS Regression, SAR Model

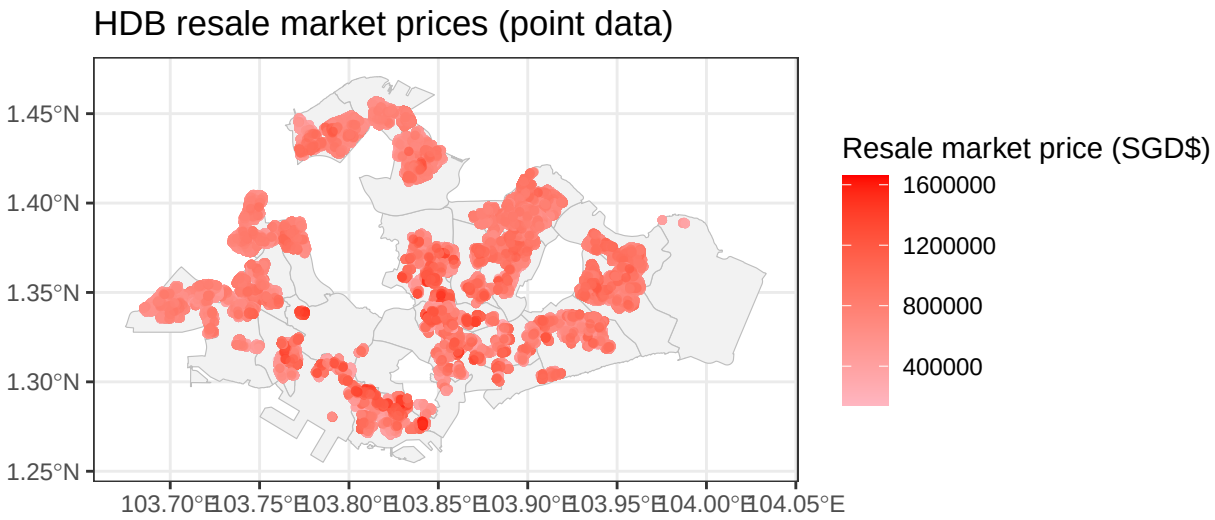
Main analysis

Data cleaning and simple descriptive analysis

The `hdb` and `resale` datasets were merged into a single dataset, `resale_sf`, which contains flat attribute data linked to geometry point data for each HDB flat. For clarity, the data cleaning code chunks to merge

datasets have been hidden from the PDF output. This dataset comprises of 218018 observations of HDB unit market resale prices between January 2017 and November 2025. We can see that we do not have resale prices across all planning areas, only in some of them.

```
# Plot geostatistical data
ggplot() +
  geom_sf(data = sg_boundary, fill = "grey95", color = "grey") +
  geom_sf(data = resale_sf, aes(color = resale_price), size = 1, alpha = 0.6) +
  scale_color_gradient(low = "lightpink", high = "red", na.value = "grey90") +
  labs(title = "HDB resale market prices (point data)",
       color = "Resale market price (SGD$)") +
  theme_bw()
```

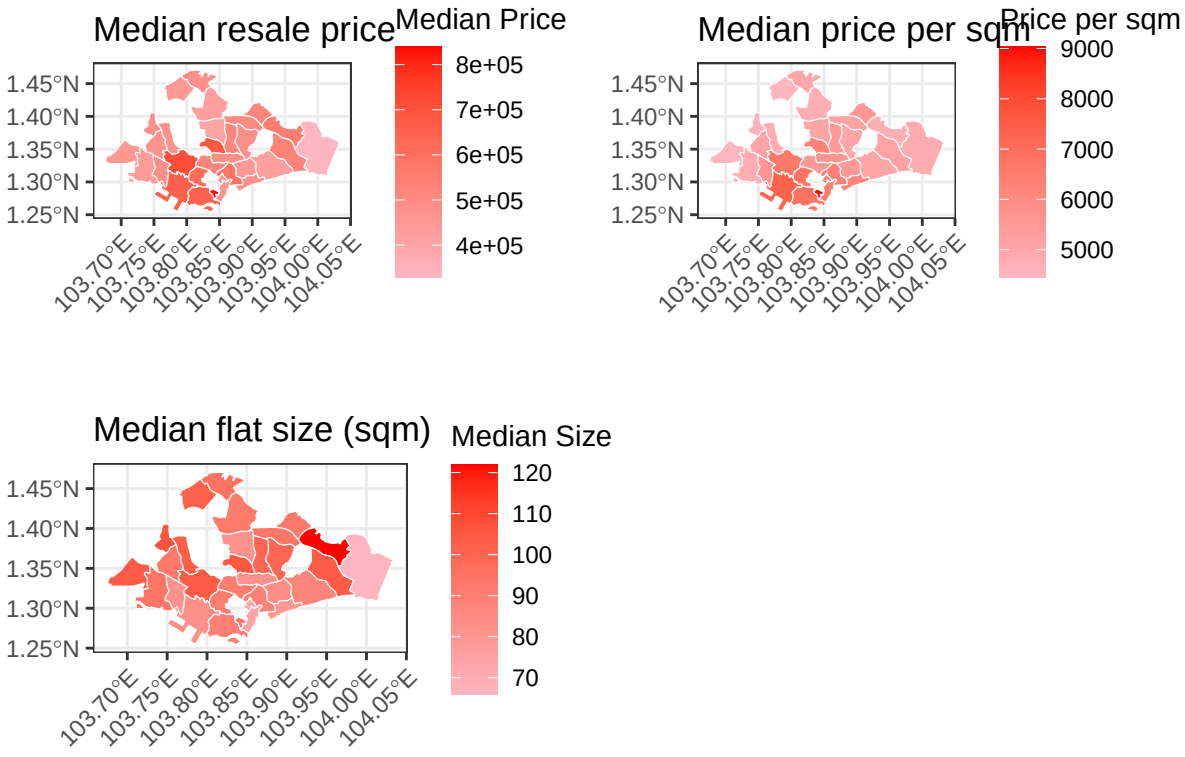


The `resale_aggregated` dataset has been created to summarise the median price, median square meters (flat size), and median price per square meter of each of the 31 planning areas with resale price data.

Visually, we observe that HDB resale prices are higher in planning areas close to the center of the map, and prices per square meter are also higher closer to the center of the map. The size by floor square meter appears to be larger when moving away from the center,

```
plot <- sg_boundary %>%
  left_join(resale_aggregated, by = "PLN_AREA_N")

p1 <- ggplot(plot) +
  geom_sf(aes(fill = med_price), color = "white") +
  scale_fill_gradient(low = "lightpink", high = "red", na.value = "grey90") +
```

```
  labs(title = "Median resale price", fill = "Median Price") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

p2 <- ggplot(plot) +
  geom_sf(aes(fill = med_psqm), color = "white") +
  scale_fill_gradient(low = "lightpink", high = "red", na.value = "grey90") +
  labs(title = "Median price per sqm", fill = "Price per sqm") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

p3 <- ggplot(plot) +
  geom_sf(aes(fill = med_sqm), color = "white") +
  scale_fill_gradient(low = "lightpink", high = "red", na.value = "grey90") +
  labs(title = "Median flat size (sqm)", fill = "Median Size") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Combine using patchwork
blank <- ggplot() + theme_void()

combined_plot <- (p1 | p2) / (p3 | blank)
combined_plot
```



Of the 55 planning areas in Singapore, 31 contained HDB resale transaction and we include these in the estimation of area-level size–price coefficients. The remaining 24 planning areas with no transactions are

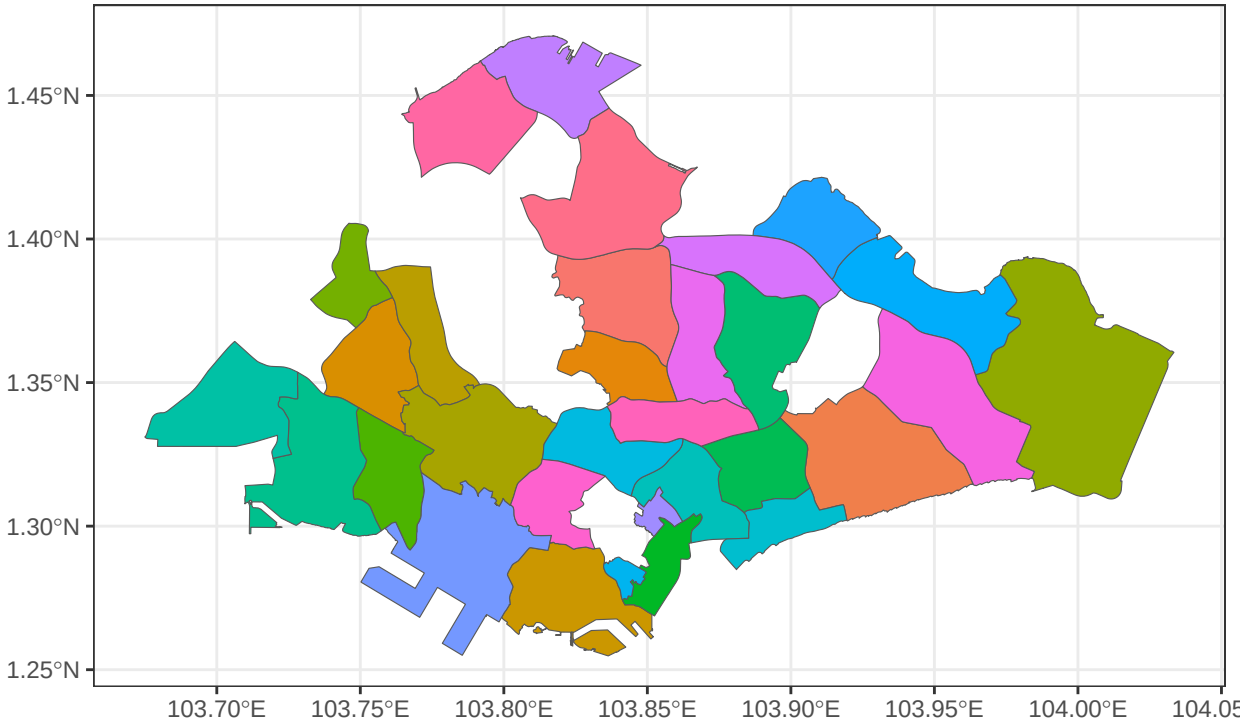
excluded from the Global and Local Moran's I calculations.

The `sg_boundary` dataset has been cleaned to contain the areal geometry data of 31 planning areas in Singapore where resale price data exists in `resale_sf`. Planning areas were selected as the unit of analysis because they provide an appropriate level of geographic granularity for comparing neighbouring districts, without being as broad and heterogeneous as regions or as highly fragmented as subzones.

```
sg_boundary %>% st_drop_geometry() %>%
  summarise(n = n_distinct(PLN_AREA_N)) %>% pull(n)
```

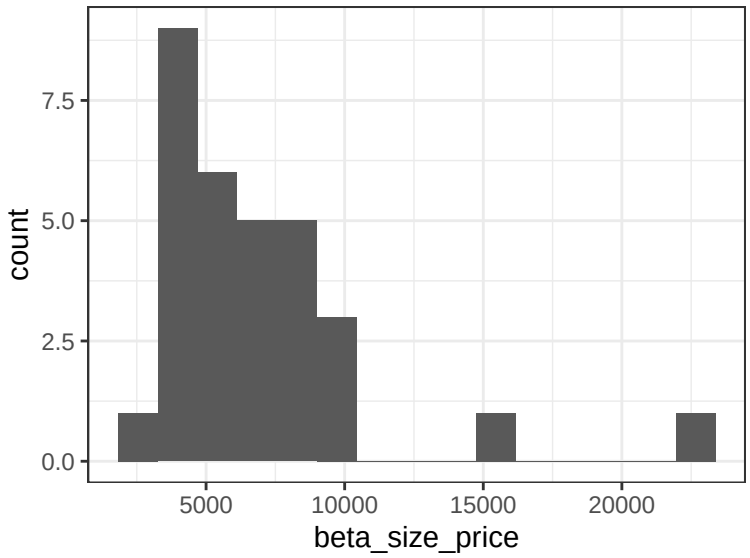
```
## [1] 31
```

```
ggplot(sg_boundary) +
  geom_sf(aes(fill = PLN_AREA_N)) +
  theme_bw() +
  guides(fill = "none")
```



To examine whether the strength of the size–price relationship exhibits spatial clustering later, we first estimated a planning-area–specific price–size coefficient (`beta_size_price`), by running a linear model resale_price ~ floor_area_sqm for each planning area. The result was saved in `sg_boundary`. The response variable exhibits non-normality, with some outliers.

```
ggplot(coeff, aes(x = beta_size_price)) +
  geom_histogram(bins = 15) +
  theme_bw()
```



Research Question 1: Is the strength of the relationship between HDB flat size and resale prices spatially clustered across Singapore's planning areas?

Assess spatial autocorrelation of the strength of the HDB size–price relationship, using Global permutation test for Moran's I on planning-area size–price coefficients.

Response Variable

- `beta_size_price`: The coefficient of the following regression was used as the size-price coefficient: lm(resale_price ~ floor_area_sqm).

Permutation test (Moran's I) was chosen as the response variable, as the normality assumption may not hold for the size-price coefficient (`beta_size_price`), and we do not have information on differences in distributions of these slope coefficients.

First, we create the neighbour list using queen-style contiguity. Queen-style is more appropriate as planning areas that are touching even at one point should be considered neighbours, given the polygons' irregular shapes.

We then created two types of spatial weights from the neighbour list: W weights (row-standardized) and B weights (binary). Using both allows us to check the robustness of our results to different weighting schemes.

```
nb_queen_2 <- poly2nb(sg_boundary, queen = TRUE)

w_weight <- nb2listw(nb_queen_2, style = "W")
b_weight <- nb2listw(nb_queen_2, style = "B")

set.seed(5226)
moran_global_w <- moran.mc(sg_boundary$beta_size_price, w_weight, nsim = 999)
set.seed(5226)
moran_global_b <- moran.mc(sg_boundary$beta_size_price, b_weight, nsim = 999)
```

```
moran_global_w
```

```
##
##  Monte-Carlo simulation of Moran I
##
## data:  sg_boundary$beta_size_price
## weights: w_weight
## number of simulations + 1: 1000
##
## statistic = 0.39876, observed rank = 997, p-value = 0.003
## alternative hypothesis: greater
```

```
moran_global_b
```

```
##
##  Monte-Carlo simulation of Moran I
##
## data:  sg_boundary$beta_size_price
## weights: b_weight
## number of simulations + 1: 1000
##
## statistic = 0.31607, observed rank = 998, p-value = 0.002
## alternative hypothesis: greater
```

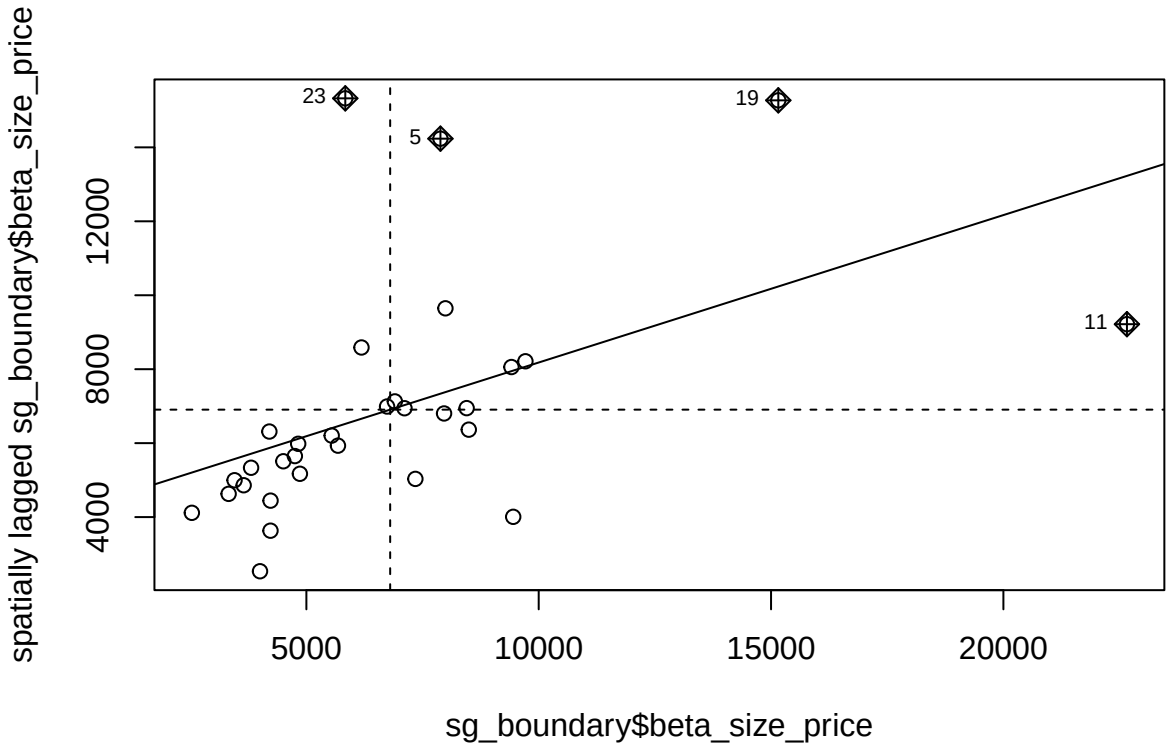
Both W and B weights produced outputs of p-values that are smaller than .05. This indicates robust evidence that the strength of the size–price relationship exhibits spatial clustering across Singapore's planning areas. i.e. planning areas with stronger price–size relationships tend to be located near areas exhibiting similarly strong relationships.

Research Question 2: Which planning areas show stronger or weaker links between flat size and resale price, compared to their neighbouring planning areas?

Detect clusters using Local Moran's I

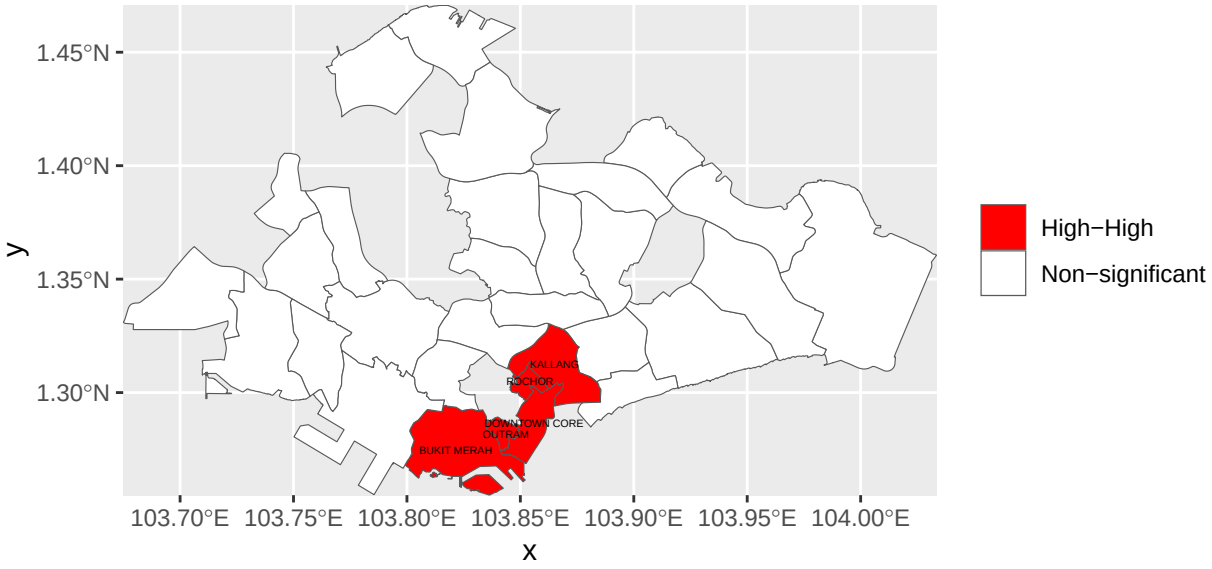
Local Moran's I (LISA) is used to identify clusters of areas with strong or weak size–price relationships relative to their neighbours.

```
loc_moran_2 <- localmoran(sg_boundary$beta_size_price, w_weight, alternative = "two.sided")

mp_2 <- moran.plot(sg_boundary$beta_size_price, w_weight)
```




```
sg_boundary$quadrant <- NA
p.2side <- loc_moran_2[, "Pr(z != E(Ii))"]

# high-high
sg_boundary[(mp_2$x >= 0 & mp_2$wx >= 0) & (p.2side <= 0.05), "quadrant"] <- 1

# low-low
sg_boundary[(mp_2$x <= 0 & mp_2$wx <= 0) & (p.2side <= 0.05), "quadrant"] <- 2

# high-low
sg_boundary[(mp_2$x >= 0 & mp_2$wx <= 0) & (p.2side <= 0.05), "quadrant"] <- 3

# low-high
sg_boundary[(mp_2$x <= 0 & mp_2$wx >= 0) & (p.2side <= 0.05), "quadrant"] <- 4

# non-significant
sg_boundary[(p.2side > 0.05), "quadrant"] <- 5

ggplot(data = sg_boundary) +
  geom_sf(aes(fill=as.factor(quadrant))) +
  geom_sf_text(
    data = subset(sg_boundary, quadrant %in% c(1,2,3,4)),
    aes(label = PLN_AREA_N),
    size = 1.5, color = "black") +

 coord_sf(expand = FALSE) +
```

```
scale_fill_manual(breaks=c(1,2,3,4,5),
labels = c("High-High", "Low-Low", "High-Low", "Low-High", "Non-significant"),
values=c("red", "blue", "lightpink", "skyblue2", "white")) +
theme(legend.title=element_blank())
```



Based on the plot, the local Moran's I identified High–High clusters of planning areas located in Central and Southern Singapore. These areas have a stronger price–size relationship (higher size–price coefficients) and are surrounded by planning areas with similarly strong relationships. There were no detectable clustering elsewhere.

Research Question 3: Amongst the identified amenity features (MRT, school, retail, park, and healthcare proximity), which are the strongest factors in explaining variation in the marginal willingness to pay for additional floor area (beta_size_price) across Singapore's planning areas?

Assess how proximity to five key amenities explains variation in the size–price gradient across planning areas using an OLS regression, followed by a Simultaneous Autoregressive (SAR) model to correct for residual spatial dependence

Data Processing Amenity Variables and Creating the 'sg_boundary_amenities' spatial object for study

To understand the possible factors underlying the differential strength of the resale size–price relationship across Singapore's planning areas, five categories of spatial amenities were incorporated into the analysis: (1) healthcare, (2) public transport, (3) public education, (4) markets and food centres, and

(5) recreational parks. The corresponding datasets used - `chas_clinics`, `mrt_exits`, `school_zones`, `market_and_food_centres`, and `parks` - act as useful approximations for the spatial presence of the abovementioned amenities. These datasets were obtained from data.gov.sg as GeoJSON files containing geospatial geometries and coordinate information. All amenity layers were imported and transformed into the same coordinate reference system (CRS) as the planning□area polygons in `sg_boundaries` to ensure spatial compatibility. As with earlier sections, the full data-processing code has been concealed from the PDF output for readability.

Centroid points were then computed for every planning area in a new spatial object, `sg_boundary_amenities`, representing the approximate centre of housing activity within each region. For 4 point-based amenities (MRT exits, CHAS clinics, markets and food centres, and parks), we calculated the minimum Euclidean distance from each centroid to the nearest amenity of the respective type. School zones, which are provided as polygonal regions, were converted into representative point locations using `st_point_on_surface()`. The resulting points, effectively the centroids of the school-zone polygons, were then used to compute the minimum Euclidean distance from each planning-area centroid to the nearest school zone. The different data processing done for the different amenities were conducted in consideration of the original data representation type in their GeoJSON files. The final dataset therefore augments each planning area with five continuous proximity indicators, all measured in metres.

All spatial analyses — including the estimation of `beta_size_price` and the Global and Local Moran's I diagnostics — were conducted on the same subset of 31 planning areas as in the earlier sections.

The resulting dataset, `sg_boundary_amenities`, contains one polygon geometry for each included planning area, together with eight attributes:

- `PLN_AREA_N` – planning-area name.
- `geometry` – planning-area polygon (sf object, EPSG:4326).
- `quadrant` – Local Moran's I cluster classification (high–high, low–low, high–low, low–high, or non-significant).

Response Variable

- `beta_size_price` – estimated marginal willingness to pay for an additional square metre, obtained from planning area-level resale regressions conducted above

Predictor Variables - i.e. Amenity Variables

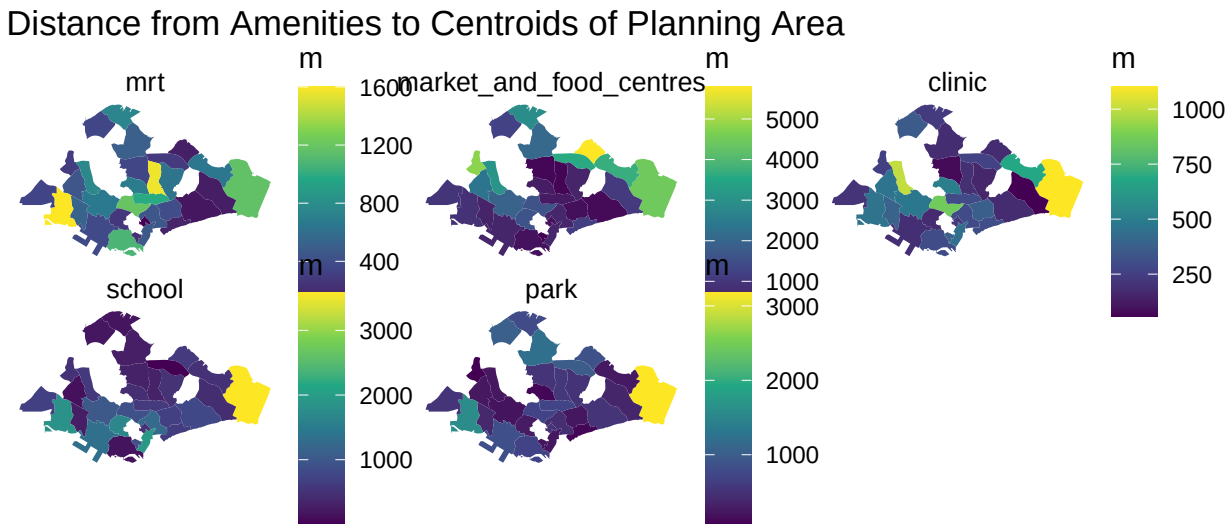
- `dist_mrt` – minimum centroid-to-MRT distance (m).
- `dist_market_and_food_centres` – minimum centroid-to-market/food centre distance (m).
- `dist_clinic` – minimum centroid-to-clinic distance (m).
- `dist_school` – minimum centroid-to-school-zone (centroid) distance (m).
- `dist_park` – minimum centroid-to-park distance (m).

All five amenity variables were computed consistently from the same set of centroids, ensuring comparability across planning areas can. Importantly, the cleaned dataset contains no missing values for any variable used in Research Question 3.

In generating the centroids for the amenities, we have assumed that the centre of the planning areas is also the centre of all housing activity within the planning area, and that housing activity is uniformly distributed within the planning area. To the best effort, the most suitable approximate data has been used to represent the locations of the amenities. In addition, for the purposes of this analysis, we assume that all amenities within each category have been accurately captured by the 5 datasets used.

As a start, the spatial patterns in amenity proximity across Singapore can be visualised in the set of choropleth maps below. The maps display minimum centroid-to-amenity distances for all five amenity types, and provide an initial overview of how the proximity and access to amenities varies regionally and forms the foundation for the analysis presented later.

```r
#Plot the distribution of all amenities by planning area
plot_list <- lapply(c("dist_mrt", "dist_market_and_food_centres",
                      "dist_clinic", "dist_school", "dist_park"),
  function(var) {
    ggplot(sg_boundary_amenities) +
      geom_sf(aes(fill = .data[[var]]), color = NA) +
      scale_fill_viridis_c() +
      labs(title = gsub("dist_", "", var), fill = "m") +
      theme_void() +
      theme(
        plot.title = element_text(size = 10, hjust = 0.5))})

wrap_plots(plot_list, ncol = 3) +
  plot_annotation(title = "Distance from Amenities to Centroids of Planning Area")
```



OLS Model Fitting

We first carry out an OLS model to act as an initial, non-spatial assessment of how proximity to key amenities relates to the strength of the resale size–price gradient across Singapore's planning areas. While the model explains a moderate proportion of the variation in `beta_size_price` ($R^2 = 0.565$), only three amenities show statistically significant associations: planning areas located further from markets/food centres and further from parks tend to exhibit lower size–price gradients, whereas areas further from schools show higher gradients. Distance to MRT exits and clinics are not statistically meaningful predictors. However, the residual spread suggests that important spatial structure is left unmodelled.

```r
# Fit OLS model
ols_model <- lm(beta_size_price ~ dist_mrt + dist_market_and_food_centres + dist_clinic +
                dist_school + dist_park,
                data = sg_boundary_amenities)

summary(ols_model)
```

```
##
## Call:
## lm(formula = beta_size_price ~ dist_mrt + dist_market_and_food_centres +
##     dist_clinic + dist_school + dist_park, data = sg_boundary_amenities)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3681.1 -1659.7  -275.9  1024.0  8625.4
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 6855.6541  1107.5366   6.190 1.79e-06 ***
## dist_mrt                      -1.0904     1.6220  -0.672 0.507573
## dist_market_and_food_centres  -0.8536     0.4016  -2.125 0.043608 *
## dist_clinic                    0.7582     2.8073   0.270 0.789306
## dist_school                    4.4285     0.9953   4.449 0.000155 ***
## dist_park                     -2.8370     1.2613  -2.249 0.033558 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2838 on 25 degrees of freedom
## Multiple R-squared:  0.565,  Adjusted R-squared:  0.478
## F-statistic: 6.494 on 5 and 25 DF,  p-value: 0.0005363
```

```r
# Extract residuals
ols_res <- residuals(ols_model)
```

Moran I's test to understand if there is spatial autocorrelation

```r
moran_ols <- moran.test(ols_res, nb2listw(nb_queen_2))
moran_ols
```

```
##
##  Moran I test under randomisation
##
## data:  ols_res
## weights: nb2listw(nb_queen_2)
##
## Moran I statistic standard deviate = 3.034, p-value = 0.001207
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##        0.35651972       -0.03333333        0.01651050
```

Because the Moran's I test on the OLS residuals revealed significant positive spatial autocorrelation (I = 0.3565, p = 0.0012), the OLS assumptions of independent errors were violated, indicating that nearby planning areas still share unexplained similarities even after controlling for amenity proximity

Modelling spatial dependence via a SAR model

To explicitly model this remaining spatial dependence, we constructed a queen-contiguity spatial weights matrix and row-standardised it (W-style) using `nb2listw()`. We then fitted a spatial error model via `spautolm()` with the same set of predictor variables. In this specification, spatial dependence is captured in the error term rather than in the outcome itself: the model assumes that unobserved factors are spatially correlated across neighbouring planning areas. Comparing the spatial error model to the baseline OLS allows us to assess whether accounting for this spatially structured error process meaningfully changes the estimated effects of amenity proximity on the size–price gradient and reduces residual spatial autocorrelation.

```r
# Spatial weights (W-style row-standardized)
lw <- nb2listw(nb_queen_2, style = "W")

# Fit Spatial Error Model
sar_model <- spautolm(
  beta_size_price ~ dist_mrt + dist_market_and_food_centres  + dist_clinic + dist_school + dist_park,
  data = sg_boundary_amenities,
  listw = lw,
  family = "SAR"
)

summary(sar_model)
```

```
##
## Call:
## spautolm(formula = beta_size_price ~ dist_mrt + dist_market_and_food_centres +
##     dist_clinic + dist_school + dist_park, data = sg_boundary_amenities,
##     listw = lw, family = "SAR")
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3255.6 -1521.1  -227.0  1153.6  7892.6
##
## Coefficients:
##                                Estimate Std. Error z value  Pr(>|z|)
## (Intercept)                  6121.48506 1100.03265  5.5648 2.624e-08
## dist_mrt                       -0.94751    1.16023 -0.8167  0.414126
## dist_market_and_food_centres   -0.67561    0.36792 -1.8363  0.066313
## dist_clinic                     1.09649    1.94904  0.5626  0.573719
## dist_school                     4.61693    0.76822  6.0099 1.856e-09
## dist_park                      -2.88063    0.99685 -2.8897  0.003856
##
## Lambda: 0.50177 LR test value: 6.4075 p-value: 0.011364
## Numerical Hessian standard error of lambda: 0.17079
##
## Log likelihood: -283.9297
## ML residual variance (sigma squared): 4849800, (sigma: 2202.2)
## Number of observations: 31
## Number of parameters estimated: 8
## AIC: 583.86
```

Results: Top Factors Affecting the Size–Price Gradient (`beta_size_price`)

The spatial error model improves upon the OLS results by explicitly accounting for spatially correlated unobserved factors ($\lambda = 0.502$, $p = 0.011$). After controlling for spatial dependence, three amenity variables emerge as the strongest predictors of the size–price gradient:

1. Distance to Schools (`dist_school`) was the strongest predictor of size–price gradient

   - Estimate = +4.62, $p < 0.001$
   - Planning areas further from school zones exhibit substantially higher beta_size_price
   - Interpretation: In areas with limited access to schools, buyers appear to value additional floor area more strongly. This may reflect that larger homes compensate for weaker school access, or that school-rich regions attract buyers for reasons less to home size.

2. Distance to Parks (`dist_park`) was also another significant negative predictor of size–price gradient

   - Estimate = –2.88, $p = 0.0039$
   - Areas further from parks exhibit lower beta_size_price.
   - Interpretation: Access to recreational space enhances willingness to pay for additional indoor space—perhaps due to lifestyle complementarities (e.g., family-oriented neighbourhoods valuing both greenery and interior space).

In summary, school proximity and park proximity are the strongest determinants of why some planning areas exhibit steeper size–price relationships than others. Other factors like Market & Food Centre proximity, MRT station proximity, and clinic proximity showed less significant association with `beta_size_price`, indicating that while these amenities may affect overall resale prices, they do not meaningfully influence the marginal value of additional floor area.

Conclusion

Across all 31 planning areas with available HDB resale transactions, our analysis reveals clear spatial structure in how strongly resale flat size influences their prices.

Research Question 1 established that the size–price gradient (`beta_size_price`) is not randomly distributed: both W- and B-style Moran's I permutation tests detected significant positive spatial autocorrelation, indicating that planning areas with strong (or weak) size–price relationships tend to cluster geographically.

Research Question 2 used Local Moran's I to map these patterns, identifying a prominent High–High cluster in the central and southern regions of Singapore, where planning areas consistently exhibit steeper size–price gradients alongside similarly strong neighbours. No Low–Low clusters were detected, suggesting that weak size–price relationships are more spatially dispersed.

Building on these spatial patterns, Research Question 3 examined whether variation in proximity to key amenities helps explain why some areas exhibit steeper size–price relationships than others. Using an OLS model followed by a SAR spatial error model to correct for residual spatial dependence, we found that two amenities: schools and parks emerged as the top two strongest predictors of `beta_size_price`. Interestingly, areas located further from school zones had systematically higher size–price gradients, while poorer access to parks was associated with lower gradients. In contrast, distances to markets/food centres, MRT stations and clinics showed much weaker meaningful influence as a predictor of `beta_size_price`. Together, these results suggest that while overall resale prices are affected by many factors, the marginal willingness to pay for additional space is most strongly shaped by school proximity and recreational environments - and that these effects operate within a spatially dependent urban structure.

Members' contributions

For Part I, Rachel prepared the solutions for Question 1 and Liying prepared the solutions for Question 2, with both members jointly discussing the questions and verifying the final answers.

For Part II, Rachel performed the primary data cleaning for the main dataset used to generate the planning-area beta_size_price coefficients and wrote the sections for Research Questions 1 and 2. Building on this foundation, Liying processed the amenity datasets for the analysis of amenity factors affecting the beta_size_price coefficients, wrote the section for Research Question 3, and drafted the concluding discussion.

Both members reviewed technical details, refined the methodology, and collaborated on the writing and editing of the final report.

AI use disclaimer and References

AI tools (ChatGPT-5 and Claude) were used to support code troubleshooting, documentation writing, and refinement of explanations. The analytical design, modelling choices, and final interpretations were fully determined by the authors. No paper and books beyond lecture notes were referenced.