# GY476 2024 SUMMATIVE ASSESSMENT

**Assessment**: The project is worth 100% of your final mark.

**Objective:**

You will take the role of a real-world geographic data scientist tasked to analyse some data from the city of Chicago in the United States. Your objective is to find useful insights for a variety of local decision-makers.

Note: It does not matter if you have never been to Chicago. The assessment will not be marked based on how much you know about Chicago but instead on how much you can show you have learned through analysing data and written work.

**Software**

- **GDS students**: You will need to carry out the assignment fully in R and compile it in R-markdown.
- **Non-GDS students**: You can choose to conduct your assessment in either QGIS or R. You can also choose to present your answers in R-markdown or use a word processor (Google Docs, Microsoft Word, etc.)

**Structure of the Report:**

- Your report should address all required points indicated in the Instructions below, so read them carefully: Introduction, questions in part A, part B and part C, and Conclusion.
- Your report should be presented as a professional piece of work. It should combine your narrative (describing your thought process, arguments, and results in English) with the figures required in the Instructions.
  - Besides the figures required, you have the option to include one more figure and one table if needed to support an argument you might want to make. Including more figures or tables than those (e.g. more than 10 figures and 1 table) will count negatively towards your mark.
  - If you use Rmarkdown, ensure that only your narrative and figures are visible when you compile your report. You can do that by adjusting the global chunk options as follows:

    ```
    knitr::opts_chunk$set(echo=FALSE, results='hide', warning=FALSE, message=FALSE)
    ```
- Your work must be fully reproducible, so the evaluators of your work can trace or replicate the steps you took to generate output. This is achieved automatically by using R and Rmarkdown (as your code summarises your steps). If you use QGIS, you would need to include a brief description of the steps and tools/packages used to generate your output as an Appendix to your report.
- Your submission should be *equivalent* to 4000 words, including code and descriptions. To ensure this requirement is met, aim for approximately 2500 words of narrative content. You do not need to count the words in the appendix or code sections separately. The package `wordcountaddin` can be used to count words in RMarkdown documents (see details in the Appendix below).

**Submission**

- You must submit one electronic copy of your summative assessment via a SharePoint link, which can be found in the Assignment Submissions folder on Moodle.
- The format of the file must be a .zip (zipped folder), including:
  - Report: containing an HTML *and* an R-markdown document, if the assessment is completed in R-markdown; or a pdf document, if another editor is used.
  - Any additional data you used in the assessment that is not specified in the instructions.
- *Please do not include your name anywhere in the documents.* If your folder paths include your name, make sure to anonymise your .Rmd file before submission.
  - For instance, instead of: ✗ `path_to_folder <- "/Users/anavarela/GY476_data/"`
  - You could write: ✓ `path_to_folder <- "/Users/XXXXXX/GY476_data/"`
  - Or: ✓ `path_to_folder <- XXXXXX`
  - Note that you should only do this right before submission, otherwise you are not going to be able to knit your file
- Name your file as follows: Course_Candidate number (e.g., GY476_34567.zip). Don't worry if your file gets renamed and please do not tell the course instructors if it does as files should remain anonymous. If there is any issue with submissions you would like to flag, please reach out to the GDS program coordinator (Hanna Wolodzko, at geog.gds@lse.ac.uk )

**Use of AI**

**Written narrative:** in this course, we follow the Department of Geography **Policy on the unauthorised use of generative AI**. On Moodle, and also below:

You confirm by submitting any assignment for GY courses that you have not used any unauthorised form of generative artificial intelligence tool in the planning and writing of any formative or summative assessment. The Department uses the following principles to distinguish between authorised and unauthorised use:

- The Department recognises that its students will want to use Generative AI to assist with ideas / understanding / essay structure, and within bounds this is acceptable.
- No AI generated text / content should be submitted as part of your assessments (whether summative or formative).

**Coding**: in this course, you can use AI for assistance with coding-related tasks. Be aware, however, that advice provided by the model may not be accurate (e.g. using Python functions as R ones) or up to date (e.g. relying on deprecated packages). It is also prone to 'hallucinate' in unpredictable ways. Exercise discretion when applying suggestions and always verify the relevance of the information in the context of the present date.

Always a good idea to follow **tips for effective use** (below and also on Moodle):

- Always verify factual information from reliable sources.
- Use AI as a starting point, not a final answer.
- When asking for feedback, request extra criticism to avoid sycophancy.
- Critically engage with the LLM through dialogue and have it take on different personas – its default first outputs are rarely very good.
- Always read original sources when summarising or referencing papers.

**Literature**

It is important to contextualise our research as social scientists to ensure that we understand the broader environment in which we carry out our analyses. In this assignment, you will need to cite at least five references that help you understand the topic from a social science perspective. Two references are provided below, and you can use them as part of your five sources. Additionally, you are encouraged to use references from the course.

When selecting sources, make sure to prioritise peer-reviewed journals and reports from reputable sources. You may also refer to technical sources (such as a blog post on formatting a map), but these are optional and do not count towards the five required references. You can use whatever citation style you prefer—just make sure to be consistent throughout. Include a bibliography at the end of your report.

References:

- Aaronson, D., Hartley, D., & Mazumder, B. (2021). The Effects of the 1930s HOLC "Redlining" Maps. *American Economic Journal: Economic Policy*, 13(4), 355–392.
- Wachsmuth, D., & Weisler, A. (2018). Airbnb and the rent gap: Gentrification through the sharing economy. *Environment & Planning A*, 50(6), 1147–1170.

**Data**

The assignment relies mainly on data from two sources, explained in more detail below.

**Airbnb listings**: Data made available on Murray Cox's website as part of his "Inside Airbnb" project which you can download here (http://insideairbnb.com/). The website periodically publishes snapshots of Airbnb listings around the world. You need to download two files from this site:

- Listings.csv, described as "Summary information and metrics for listings in Chicago (good for visualisations)"
- neighbourhoods.geojson, "GeoJSON file of neighbourhoods of the city."

**Socio-economic variables**. Source: American Community Survey (ACS) 2018-2022, U.S. Census Bureau.

- A subset of variables from the latest ACS has already been retrieved for you in ACS_2018_2022_cook_vars.csv, and is available in Moodle. This includes data for Cook county (Illinois), that contains Chicago.
- A geopackage with the Cook county census tracts has also been made available to you on Moodle.
- For more information about the ACS (2018-2022) you can have a look at:
  - https://www.census.gov/data/developers/data-sets/acs-5year.html
  - https://api.census.gov/data/2022/acs/acs5/variables.html
  - Cook County: https://data.census.gov/table?g=050XX00US17031$1400000&y=2022&d=ACS%205-Year%20Estimates%20Detailed%20Tables

## INSTRUCTIONS

**INTRODUCTION:** Include a brief introduction to your report.

## PART A: COMPLETE ALL QUESTIONS

### 1. Description of data and context

Load the Airbnb listings data into your working environment, and explore it. In your assessment submission, describe:

- The two main datasets used for this project (Airbnb listings and ACS): sources, strengths and limitations of these types of data.
- You might consider dropping some of the Airbnb listings outliers for better visualising some results (e.g. drop from your analysis those listings with the highest prices). In your answer, describe whether you have decided to drop some of the outliers and your reasoning.
- The CRS you are going to use in this project. Justify your answer.

### 2. Mapping and Data Visualisation: Airbnb presence at the census tract level

Summarise the data by producing the following figures:

- **Figure 1**: This figure should contain the following maps: (1) location of Airbnb listings with a room_type equal to "Entire home/apt"; and (2) location of Airbnb listings with room_type equal to "Hotel room", or "Private room", or "Shared room".
- **Figure 2**: Number of listings per census tract. Explore the spatial distribution of the data using choropleths.
- **Figure 3**: Average price per census tract. Explore the spatial distribution of the data using choropleths.

What are some key takeaways of the spatial distribution of Airbnb in Chicago that we learn from these three maps? Justify your data classification methods and visualisation choices.

### 3. Mapping and Data Visualisation: Socio-economic variables from the ACS data

Select two variables from the American Community Survey data.

**GDS students**: You can use at most one of the variables provided to you in the .csv file on Moodle, but you will have to select at least one new variable from the American Community Survey (ACS) 2018-2022 through the R package Tidycensus.
**Non-GDS students**: You can select both variables from the variables provided to you in the csv file, or you can select others from the ACS through the R package Tidycensus.

- **Figure 4**: This map should be just one figure, and contain two submaps: one choropleth for each variable selected. If you choose to calculate population percentages, make sure you standardise the table by the relevant population size of each census tract.

What can you say about the spatial distribution of your socio-economic variables of interest? Justify your data classification methods and visualization choices.

### 4. Combining Datasets

- **Figure 5**: Plot the natural logarithm of price (ln of price) of Airbnb listings in Chicago together with one of your chosen socio-economic variables of interest. There are several ways of doing this.

Comment on the details of your map and analyse the results. What does this map tell you about the relationship between Airbnb location/price and your socio-economic variable of choice?

### 5. Query OpenStreetMap data

Choose an amenity to query in OpenStreetMap (for example bars, restaurants, or public transit stations). Query your amenity of choice and save the data.

- **Figure 6**: Create a heatmap of your amenity of choice.
- **Figure 7**: Find out which Airbnb listings are within 200 metres (or less) of your amenity of choice.

Comment on the details of your maps and analyse the results.

---

## PART B: COMPLETE ONLY ONE OF THE OPTIONS BELOW

You need to pick one of the four options described below. You must include one map (**Figure 8**) to support your analysis.

For your chosen option, you need to address the following points:

- Describe the analysis type that you have decided to implement, and discuss why. Discuss as well any key parameters of your analysis that you had to decide upon.
- Visualise the output of your analysis. You might choose to incorporate some of the data available to you (e.g. Airbnb or ACS data) to communicate your message.
- Analyse your results: what have you learned from your analysis?
- What is one piece of advice you derive from your findings that might provide insights to the city decision-makers?

**Option 1: Interpolation (IDW) or Point Pattern analysis of Airbnb or OSM data**
You can choose from any of the Interpolation of Point Pattern analyses we have covered in the course. Choose which variable to focus on. If you use OSM data, it must be different from the amenity chosen in 6.

**Option 2: Network Analysis or Routing**
You can perform a type of network analysis of your choice from those covered in the course. Some data you might want to use for this (although note that this might not be needed, if you use data from the osrm package in R, for instance):

- Chicago street network: https://data.cityofchicago.org/Transportation/transportation/pr57-gg9e/about_data

- Taxi trips (this is a very large dataset, so you can select a subset of it): https://data.cityofchicago.org/Transportation/Taxi-Trips-2024-/ajtu-isnz/about_data

**Option 3: Work with Landsat data (raster) Land use classification**
You can generate a map of a NDVI index, or carry out a supervised classification algorithm similar to the one we did for London (classifying data into developed, not developed and water.) If you decide to do the latter, you should include data on the training points you generated for your classification in your submission. You are provided a Landsat image for this analysis, but can choose a different one.

**Option 4: Plotting relationships between spatial variables**
For this option, you will be using the Inside Airbnb price data together with four socio-economic variables of your choice. These can include the two you have focused on part A.
In terms of visualisation, create four figures, each plotting the average listing price and one of your socio-economic variables of interest. These figures could be scatterplots, or other types of visualizations useful to plot spatial correlations (e.g. like bivariate maps, see resources below). You can also choose to include a cross-correlation matrix of your variables.

---

## PART C: COMPLETE ONLY ONE OF THE OPTIONS BELOW

You need to pick one of the two options described below. You must include one figure (**Figure 9**) to support your analysis.

**Option 1**: **Complete another of the options under Part B above.**

**Option 2**: **Choose your own data and analysis.**
For this option, you need to find another dataset relevant for Chicago, beyond those provided in this assignment.

In your answer:

- Justify why you have chosen to focus on those data.
- Visualise your data and/or the output of any analyses you have carried out.
- Analyse your results: what have you learned from your analysis of these data?
- What is one piece of advice you derive from your findings that might provide insights to the city decision-makers?

Some potential sources of data:

- Chicago Data Portal: https://data.cityofchicago.org/
- NASA Earth Observation Data: https://earthdata.nasa.gov/

---

**CONCLUSIONS:** Include a brief conclusion statement to your report.

## APPENDIX

### Count words in Rmarkdown (excluding code chunks and inline code)

```
if(!require("wordcountaddin")) remotes::install_github("benmarwick/wordcountaddin")

word_count()
```

**Resources to help you**. See also suggested resources in slides throughout the course.

Airbnb:

- https://vizual-statistix.tumblr.com/post/114850050736/i-find-the-spread-of-airbnb-to-be- as-fascinating
- https://carto.com/blog/airbnb-impact/

US census geographies:

- https://learn.arcgis.com/en/related-concepts/united-states-census-geography.htm

Bivariate maps:

- https://cran.r-project.org/web/packages/biscale/vignettes/biscale.html
- https://bnhr.xyz/2019/09/15/bivariate-choropleths-in-qgis.html

Tidyversus package:

- https://walker-data.com/tidycensus
- https://walker-data.com/isds-webinar/#1

**Extract of the American Community Survey (ACS) 2018-2022, US Census Bureau.**
Observations: 1332; Variables: 423; Years: 2018-2022; Geographical Unit: census tract

For the list of variables in each table, check here:
https://api.census.gov/data/2022/acs/acs5/variables.html

| Table | Description |
|---|---|
| B19013_001 | Median household income in the past 12 months |
| B02001 | Population by race |
| B23006 | Population by education |
| C27012 | Population by Health insurance |
| B08006 | Commuting variable |
| B09010 | Supplementary income variables |
| B09019 | Household type counts |
| B17001 | Poverty Status |

If you want to visualise some aspects of different Subnational Administrative boundaries, you can download administrative boundaries from the U.S. Census.