Part 1

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Moran’s I** | **z-score** | **p-value** |
| **Inverse Distance** | 0.008441 | 3.709437 | 0.000208 |
| **Inverse Distance Squared** | 0.007451 | 1.453946 | 0.145961 |
| **Fixed Distance Band** | 0.009971 | 4.955611 | 0.000001 |
| **Zone of Indifference** | 0.009968 | 4.956624 | 0.000001 |

Question 1

Moran’s I measures the amount of spatial autocorrelation that there is within given variables. The four methods utilized above are used to provide variance to the weight of the data as it’s being tested. The inverse distance, fixed distance band, and zone of indifference all have pretty high z-scores and extremely low p-values. This trend means that there is clustering present. This makes sense as the weights are distributed wider and the calculations work with more of the neighbors of a particular point. Inverse distance square works similarly to that of inverse distance, however the slope is steeper and the weighting only encompasses the most-immediate points of each point. Given the high z-score and the high p-value, we can identify that there is a clustering pattern occurring but cannot reject the null hypothesis that complete spatial randomness is causing the data seen. For the tornado fatalities and injuries, the fixed distance band method is optimal because it works best with point data and with varying sizes of polygons (our counties match this well), which is what we’re working with.

The numbers run with these methods are all relatively similar—the z-scores are all high and except in the case of the inverse distance squared the p-values are relatively and consistently low. This leads me to believe that tornado fatality rates exhibit spatial autocorrelation.

Question 2

The spread of annual precipitation data across Colorado is fairly dispersed throughout the state; however, there appear to be lighter values (corresponding to less rainfall) along the range of the Rocky Mountains. There also looks to be a cluster of higher values within the Denver metro area. The lowest values look to be within the mountains and the higher values look like they’re in the southeast corner and on the western side of the state. There appears to be high variability going from east to west (the north-south values are relatively consistent) with a stark shift from low to high values where the mountains are, particularly just west of Denver.

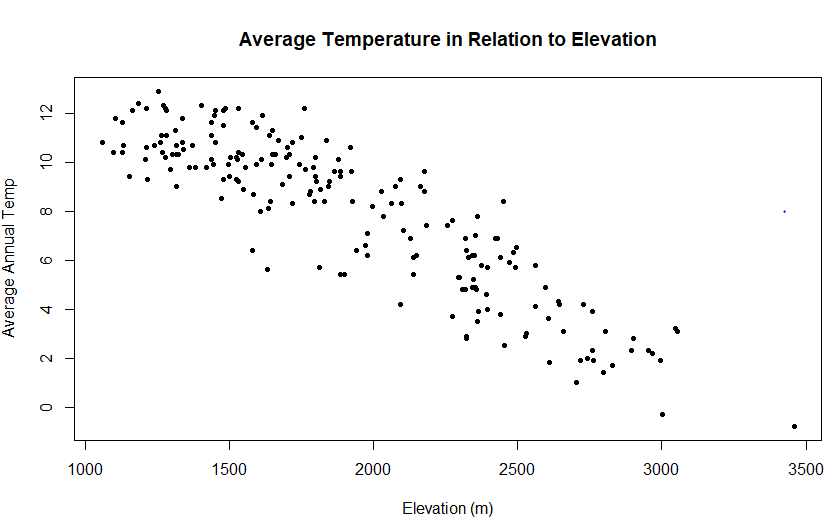
Question 3

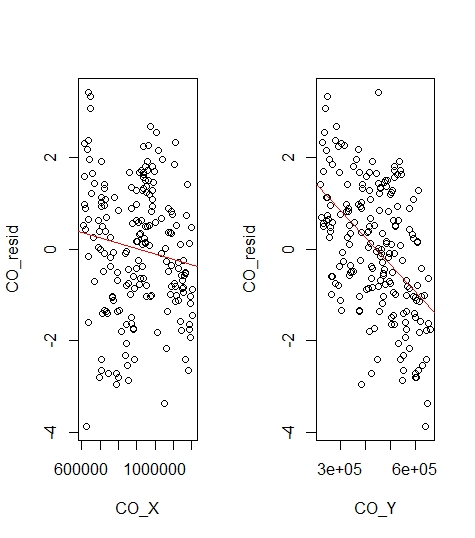
The range of the data appears to be at about a distance of 75,000m or 750km. This means that at about this distance is where the data starts to level off and we stop seeing the effects of spatial autocorrelation. In plain English, this means that the range notates about where the effects of the assumed phenomena stop affecting and/or adjusting the data to follow a particular pattern.

Question 4

The average temperate and average precipitation plots look extremely similar. They both have a range of about 750-800km that is followed by fairly inconsistent values that are on a general increasing trend. The precipitation differs slightly in that the distance values after the range have a lot more peaks and valleys, meaning that the data varies a bit more.

Question 5



In the linear regression modeled after the average annual temperature in relation to the elevation of Colorado, we find that as elevation increases, temperature decreases. This can be seen in the scatterplot above, with the general slope of the data at a steady decline. The p-value for this data is about 0, indicating that we can reject the null hypothesis that says that there is no association between elevation and average temperature. From this we can conclude that there is a correlation between these two variables. The coefficient of determination shows how much the variance in y, in this case the average annual temperature, can be described by the variance in x, or elevation. With an r-squared value of 0.7973 tells us that 79.93% of the variation in temperature is described by the variation in elevation.

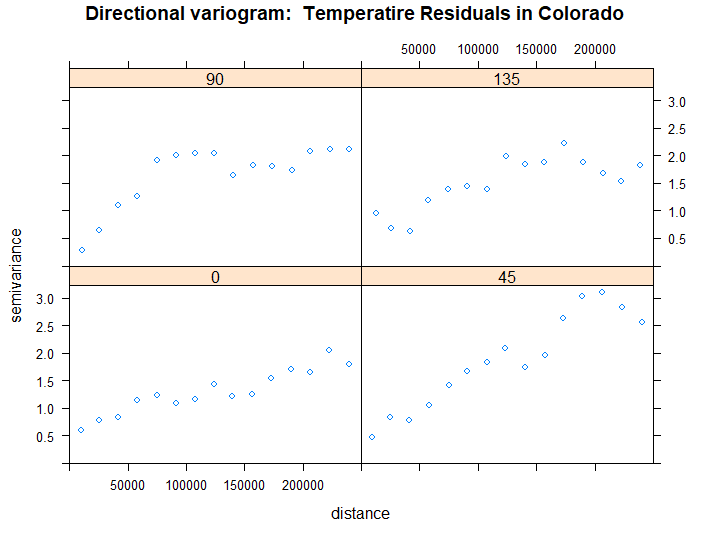
Question 6

While both of the slopes for latitude and longitude are negative and on a decline, the slope of the latitude linear regression line is steeper than that of the longitude. This means that the slope for the residuals is larger and that the expected values for latitude deviate further than that of the longitude’s given our linear model of elevation. This says that the variation, or error, of latitude is greater than the error of longitude and elevation does not predict these values as well.

Question 7

In the semivariogram of the residual values of temperature, the visual looks quite similar to the semivariogram for temperature. The biggest difference that can be seen is in the range: with the normal temperature graph, the range was about 750-800km; in this residual graph, the range has raised to be about 1100km. There are also only a few peaks and valleys instead of these values being all over the place, meaning that the numbers are much more gradual as they transition from low to high values.

Question 8



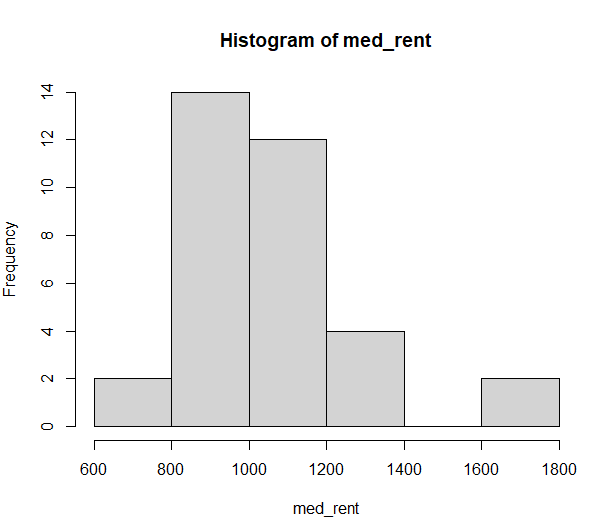
The 0 and 90 labels indicate the north and south direction; the 135 and 45 labels indicate east and west. The range of semivariance is greatest within the east. This makes sense as the temperatures and climate within the eastern part of the state has high variability, particularly since this area is full of plains and lacks mountain or varying elevations like the rest of the state. The south and west directions appear to have very similar visuals with a pretty similar range in semivariance. This also tracks as the southern and western parts of Colorado have similar climates and temperatures, as was seen within the temperature plots created earlier in the lab. After normalizing for the elevation, there appears to be a higher spatial variation amongst the east-west data than the north-south data. The north-south range is about 75km while the east-west is around 125km. The east-west trend of greater variation can be explained because we are conscious there are mountains there that affect the data and overall weather patterns in Colorado. This data therefore appears to be anisotropic, or direction-specific, since the patterns have greater impact on the data visualization in the east-west direction. The global variance appears to be different for all of them given the differing sill values.

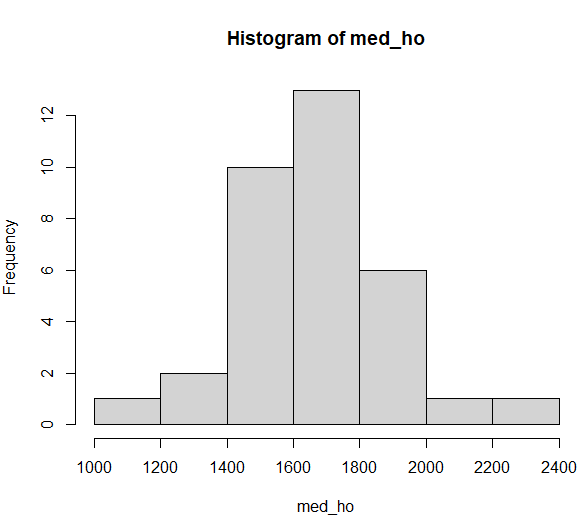
**Part 3: Characterizing Data**

The data that I used for part 3 is housing census data for the city of Bellingham, Washington. The census tract data came from the Whatcom County website:

<https://www.whatcomcounty.us/716/Data>

which is the county in which the city resides. These data were collected in 2018 as part of an American Community Survey (ACS) that was focused on collecting demographic information of the area. These surveys are given to residents throughout the United States periodically in some non-census years to keep data up-to-date. In this area, the survey was given either by mail with an option for residents to return their responses via mail or online. The ACS that I pulled data from was focused on housing trends in Bellingham and Whatcom County. The output table included fields such as median monthly rent and median monthly homeowner costs.

I ran a simple histogram of the median monthly rent column (med\_rent). The summary indicated a minimum value of $656; a maximum of $1,784; a median of $1,009.50; and a mean of $1,066.90. I decided to create a histogram so that I could try and gauge what kind of distribution and normalcy curve the data would follow. In the resulting graphic, seen below, I found that the data followed a pretty normal bell-curve with a small hiccup on the upper-range where the $1600-$1800 price point pops up a few occurrences. Overall, the median rental rates in Bellingham are skewed to the left, meaning that the aforementioned higher price ranges are borderline outliers within this dataset.

I wanted to compare the rental rates with the monthly cost of homeownership (med\_ho), so next I plugged this field into the summary and histogram functions in R. The summary told me that owning a home in the city is more expensive than renting with a range of $1,174 - $2,269 and a mean of $1,659. These values are a lot different than those within the renting column. In looking at the histogram of the data, the homeowner column shows a nearly perfectly-centered normal curve with a very slight left skew, but with a range that is consistent and without holes as the rental historigram showed. Both of these datasets would be great to use in statistical analysis as the bulk of their data falls within the primary hump of the curve of the data.

**Part 4: The Final Checkpoint**

Final Project Outline

Introduction

* Describe study area (LoDo) and contextualize the historic neighborhood within Denver
* Pose research question: how is the size of newer buildings in LoDo determined by the year that they are built?
* Discuss literature: urban planning, architecture, and urban geography – talk about building placement in urban neighborhoods and construction processes that help determine lot/building size

Methods

* Gather data and combine into a CSV; organize in Excel and upload into R.
* Run linear regression model and, if applicable, a T-test to look into potential clustering trends.
* How does this help answer the question? 🡪 shows correlation between the two variables and compares their spatial distribution within the study area

Results

* What did I find? 🡪 Analyze and summarize the outcomes of the tests in plain English—e.g. explain what the numbers means and describe the correlations identified
* Include plotted linear regression line as well as t-test output table for visual guidance

Limitations

* Refer back to literature review—what limitations were there in those studies? Are those limitations present in my study?
* Limitation of data collection: human error as I learn to become more comfortable with county datasets and performing statistical tests within R Studio

Discussion

* Main findings 🡪 refer back to the ‘so what?’ and discuss why the findings are important.
* What does the research imply for other urban neighborhoods? If there is a correlation between size and building year, what does this mean for neighborhoods within downtown areas? Can we expect cities to get ‘bigger’ if this is the case? Be sure to reference statistics to back up claims
* If there is clustering present, what could this mean for the future of the LoDo neighborhood?
* Next steps: contribute further to the discipline so that urban planners can adequately gauge how urban neighborhoods are changing and adjust city plans to accommodate minority populations within downtown-city-limits.