**Larger AUC= higher drug sensitivity**

**Smaller IC50= higher drug sensitivity**

Introduction questions:

What is correlation?

Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate). It’s a common tool for describing simple relationships without making a statement about cause and effect. (<https://www.jmp.com/en_ca/statistics-knowledge-portal/what-is-correlation.html>)

Correlation is the relationship between two variables measured with the correlation coefficient. This coefficient ranges from -1 to +1. A positive correlation associates higher values of a variable with higher values of another variable. A negative correlation associates higher values of a variable with lower values of another variable.

What is the difference between Pearson, Spearman, and Matthews correlation?

1. **Pearson Correlation Coefficient**: This measures the linear relationship between two variables.

2. **Spearman Correlation Coefficient**: This measures the rank-order relationship (monotonic relationship) between two variables.

3. **Matthews Correlation Coefficient**: is a statistical tool used for model evaluation. Its job is to gauge or measure the difference between the predicted values and actual values and is equivalent to chi-square statistics for a 2 x 2 contingency table.

**Question 1: Create a scatterplot for each drug comparing the AUC in GDSC and CCLE for all cell lines (hint: code from Tutorial 1b may help).**

| summarizedData |>  ggplot(aes(x = auc\_GDSC, y = auc\_CCLE)) +   geom\_point(alpha = 1/2,) +  facet\_wrap(~ drug) +  xlab("GDSC AUC") +  ylab("CCLE AUC") +  ggtitle("AUC summaries of cell line response to AZD0530 across studies") |
| --- |



**Question 2:** Calculate the Pearson and Spearman correlation coefficients of the AUC in GDSC and CCLE for each drug (hint: code from Tutorial 1b may help).

| # A tibble: 15 × 3  drug Pearson\_auc Spearman\_auc  <chr> <dbl> <dbl>  1 17-AAG 0.555 0.569   2 AZD0530 0.515 0.444   3 AZD6244 0.412 0.290   4 Crizotinib 0.575 0.167   5 Erlotinib 0.347 0.452   6 Nilotinib 0.772 0.147   7 Nutlin-3 0.337 0.269   8 PD-0325901 0.653 0.607   9 PD-0332991 0.340 0.332  10 PHA-665752 0.164 -0.00138 11 PLX4720 0.628 0.377  12 Sorafenib 0.300 0.289  13 TAE684 0.457 0.419  14 lapatinib 0.593 0.429  15 paclitaxel 0.392 0.384 |
| --- |

**Question 3: Based on the scatterplot and correlation values for AUC, would you say that GDSC and CCLE tend to agree? Why or why not?**

ChatGPT answer:

1. **Pearson Correlation Coefficient**: This measures the linear relationship between two variables.

2. **Spearman Correlation Coefficient**: This measures the rank-order relationship (monotonic relationship) between two variables.

Yes, when we look at the scatter plot the dots are in the same direction. If we were to draw a line with the same tendency as the dots, these will be similarly close to that line.

When we look at all the values of both drugs, take their average, and compare them, we can conclude that there is a moderate relationship between the values and they tend to agree. However, when looking at each sample in detail, we observe discrepancies in the correlation between the drugs. This leads us to understand that we cannot compare the samples as a whole; rather, each one must be examined individually. Through this analysis, we can conclude that in most samples, there is not a valid correlation between both studies.

If we were to use the Pearson Correlation Coefficient average between the drugs (0.469), it would point towards a moderate linear relationship between GDSC and CCLE.

The 2 samples t-test ran got a p-value of: 0.0216. Which indicated a significant difference between the values of GDSC and CCLE.

I told chat GPT to carry out a t-test with the values; here’s what it said:

The paired t-test results for comparing the variables X and Y are as follows:

* **t-statistic**: 2.598
* **p-value**: 0.0216 (en excel)

### **Interpretation**

The paired t-test is used to compare the means of two related groups to determine if there is a statistically significant difference between these means. The null hypothesis (H0) of the paired t-test is that the mean difference between the paired observations is zero.

* If the p-value is less than the chosen alpha level (commonly 0.05), we reject the null hypothesis, suggesting that there is a statistically significant difference between the means of the two variables.
* If the p-value is greater than the chosen alpha level, we fail to reject the null hypothesis, suggesting that there is no statistically significant difference between the means of the two variables.

In this case:

* The p-value is 0.0211, which is less than 0.05.

### **Conclusion**

We reject the null hypothesis and conclude that there is a **statistically significant difference between the means of the variables X and Y** This suggests that GDSC and CCLE do not agree perfectly in their AUC values, indicating some level of disagreement between the two datasets. ​

drug pearson\_cor spearman\_cor

Length:15 Min. :0.1600 Min. :0.0000

Class :character 1st Qu.:0.3450 1st Qu.:0.2800

Mode :character Median :0.4600 Median :0.3800

Mean :0.4687 Mean :0.3453

3rd Qu.:0.5800 3rd Qu.:0.4350

Max. :0.7700 Max. :0.6100

### Analysis of Correlation Values

* **Strong Positive Correlation**: Typically, values above 0.7 are considered strong.
* **Moderate Positive Correlation**: Values between 0.3 and 0.7 indicate a moderate relationship.
* **Weak or No Correlation**: Values below 0.3 indicate weak or no relationship.

### Drugs with Strong Correlation

**Nilotinib**:

* Pearson: 0.772 (Strong)
* Spearman: 0.147 (Weak)

**PD-0325901**:

* Pearson: 0.653 (Moderate)
* Spearman: 0.607 (Moderate)

### Drugs with Moderate Correlation

- \*\*17-AAG\*\*:

- Pearson: 0.555

- Spearman: 0.569

- \*\*Crizotinib\*\*:

- Pearson: 0.575

- Spearman: 0.167

- \*\*lapatinib\*\*:

- Pearson: 0.593

- Spearman: 0.429

- \*\*PLX4720\*\*:

- Pearson: 0.628

- Spearman: 0.377

### Drugs with Weak Correlation

- \*\*PHA-665752\*\*:

- Pearson: 0.164

- Spearman: -0.00138

### Visual Assessment (Scatterplot)

To confirm these observations, it would be helpful to create scatterplots of AUC values from GDSC vs. CCLE for each drug. The scatterplot can show if there is a clear linear or monotonic trend.

Austen - in my version I made a barplot. Can redo to a scatterplot later if we prefer that:

**```{R}**

**correlation\_results <- summarizedData %>%**

**group\_by(drug) %>%**

**summarize(**

**pearson\_cor = round(cor(auc\_CCLE, auc\_GDSC, method = "pearson"), 2),**

**spearman\_cor = round(cor(auc\_CCLE, auc\_GDSC, method = "spearman"), 2)**

**)**

**# View(correlation\_results)**

**# This probably isn't necessary. I'm just obsessive and like nice graphs**

**corr\_results\_pivot <- correlation\_results %>%**

**pivot\_longer(**

**cols = c(pearson\_cor,spearman\_cor),**

**names\_to = "method",**

**values\_to = "correlation"**

**)**

**corr\_results\_pivot %>%**

**ggplot(aes(x = drug, y = correlation, fill = method, group = method)) +**

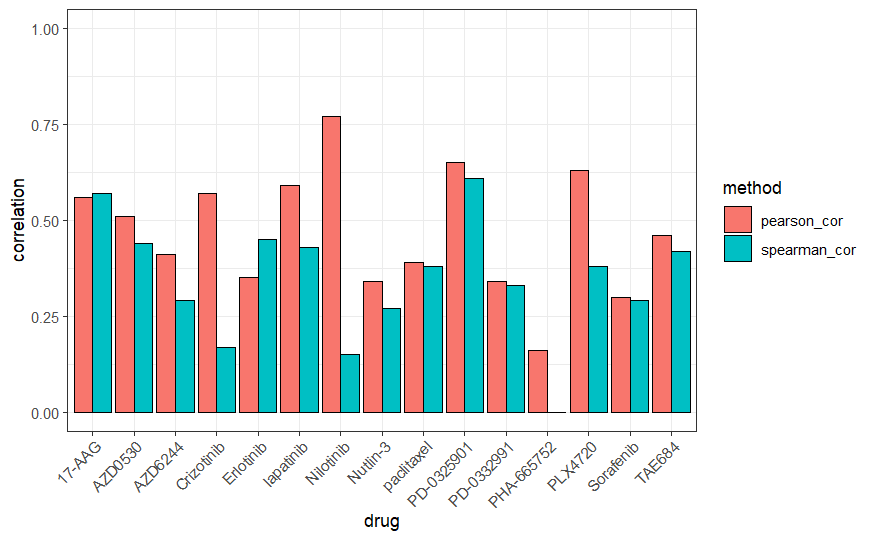
**geom\_bar(stat = "identity", position = position\_dodge()) +**

**theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +**

**ylim(0, 1) +**

**ggtitle("AUC Correlation Coefficients")**

**```**



### Agreement Between GDSC and CCLE

Based on the correlation values alone:

- There is moderate agreement for some drugs (e.g., PD-0325901, 17-AAG).

- For other drugs, the correlation is weak (e.g., PHA-665752).

### Differences Between Pearson and Spearman

- For drugs like \*\*Nilotinib\*\*, the Pearson correlation is strong, but the Spearman correlation is weak. This suggests a linear relationship that may not hold for all ranks.

- Generally, Pearson and Spearman coefficients are moderately consistent, suggesting some level of agreement, but not very strong.

### Conclusion

Overall, while there is some agreement between GDSC and CCLE, it is not very strong for all drugs. The moderate Pearson and Spearman correlations for several drugs suggest that the two studies do not consistently agree.

### Pearson vs. Spearman

- \*\*Pearson\*\* is better for detecting linear relationships.

- \*\*Spearman\*\* is better for detecting monotonic relationships.

Given the moderate correlations, neither measure alone is definitively superior for summarizing the consistency. Both provide valuable insights, and together they suggest that the agreement is present but not strong.

### Additional Measures

- To further explore the consistency, you could consider distance correlation, which captures both linear and nonlinear relationships, offering a more comprehensive view of the relationship between GDSC and CCLE.

**Question 4: Do the Pearson and Spearman correlation coefficients that you calculated differ? Which do you think is a better summary of the consistency between the two studies?**

Yes, the Pearson and Spearman correlation coefficients we calculated differ. Values range from 0.001375766 to 0.7720569.

| The Pearson and Spearman correlation coefficients do indeed differ, as they measure different aspects of the relationship between two variables:  1. \*\*Pearson Correlation Coefficient\*\*: This measures the linear relationship between two variables. For our data, the Pearson correlation coefficient between Pearson\_auc and Spearman\_auc is approximately 0.336. This indicates a moderate positive linear relationship.  2. \*\*Spearman Correlation Coefficient\*\*: This measures the rank-order (monotonic) relationship between two variables. It is not affected by the actual values but rather by their ranks. The Spearman correlation coefficient often captures nonlinear relationships better than the Pearson correlation.  ### Better Summary of Consistency  Given the context of this analysis—comparing performance metrics (Pearson\_auc and Spearman\_auc) which likely derive from different evaluation criteria—the \*\*Spearman correlation coefficient\*\* might be a better summary of the consistency between the two studies. This is because:  - \*\*Rank-based Comparison\*\*: Spearman correlation focuses on the ranks of the values, which can be more informative when comparing performance metrics that might not be linearly related.  - \*\*Robust to Outliers\*\*: Spearman is less sensitive to outliers and non-linear relationships, making it more robust in cases where the relationship between metrics is not strictly linear.  ### Consistency between the Two Studies  The relatively low Pearson correlation (0.336) suggests that the relationship between the two sets of metrics (Pearson\_auc and Spearman\_auc) is not strictly linear, implying that high values in one metric do not consistently correspond to high values in the other. The same drugs can have notably different ranks across these metrics (e.g., Nilotinib and Crizotinib), which further supports the choice of Spearman for understanding consistency in ranking rather than exact values.  Therefore, while the Pearson correlation provides useful information about the linear relationship, the Spearman correlation gives a better picture of the consistency in rankings between the two metrics, which seems more appropriate for this type of data comparison. |
| --- |

**Question 5:**

We have explored Pearson and Spearman correlation, but what about other correlation measures? For example, you could try out distance correlation, which is sensitive to nonlinear relationships. You can find this measure in the `energy` R package, which you'll need to install and load with the following commands:

Austen -

| **```{r} install.packages("energy") library(energy)   # Did this wrong the first time. dcor() is just cor() with a fancy hat distance\_correlation <- summarizedData %>%  group\_by(drug) %>%  summarize(dcor = round(dcor(auc\_CCLE, auc\_GDSC), 2))  #View(distance\_correlation)   ```** |
| --- |

The version above should work fine to make a data frame with distance correlation, but I got curious about what I could do and got a little carried away making a fancy graph…

**```{r}**

**# install.packages("energy")**

**# library(energy)**

**# Did this wrong the first time. dcor() is just cor() with a fancy hat**

**distance\_correlation <- summarizedData %>%**

**group\_by(drug) %>%**

**summarize(dcor = round(dcor(auc\_CCLE, auc\_GDSC), 2))**

**# View(distance\_correlation)**

**# Added the measures together in the data frame**

**# correlation results is just the table with the pearson and spearman**

**all\_cors <- correlation\_results %>%**

**mutate(distance\_cor = distance\_correlation$dcor)**

**# Need it to be pivoted tidy for the graph to be pretty**

**# The graph WILL be pretty**

**all\_cors\_tidy <- all\_cors %>%**

**pivot\_longer(**

**cols = c(pearson\_cor,spearman\_cor, distance\_cor),**

**names\_to = "method",**

**values\_to = "correlation"**

**)**

**# Omg I bet ChatGPT knows how to put it in descending order.**

**# This graph is gonna be so gud**

**# Calculate the average correlation for each drug**

**average\_correlation <- all\_cors\_tidy %>%**

**group\_by(drug) %>%**

**summarize(avg\_correlation = mean(correlation, na.rm = TRUE))**

**# Reorder the drug factor levels based on average correlation in descending order**

**all\_cors\_tidy <- all\_cors\_tidy %>%**

**mutate(drug = factor(drug, levels = average\_correlation$drug[order(-average\_correlation$avg\_correlation)]))**

**# Thank you Skynet. Please don't murder all humans.**

**# Fixed to show all correlation measures together.**

**all\_cors\_tidy %>%**

**ggplot(aes(x = drug, y = correlation, fill = method, group = method)) +**

**geom\_bar(stat = "identity", position = position\_dodge()) +**

**scale\_fill\_brewer(palette = "Set1") + # Changing the color palette is critical**

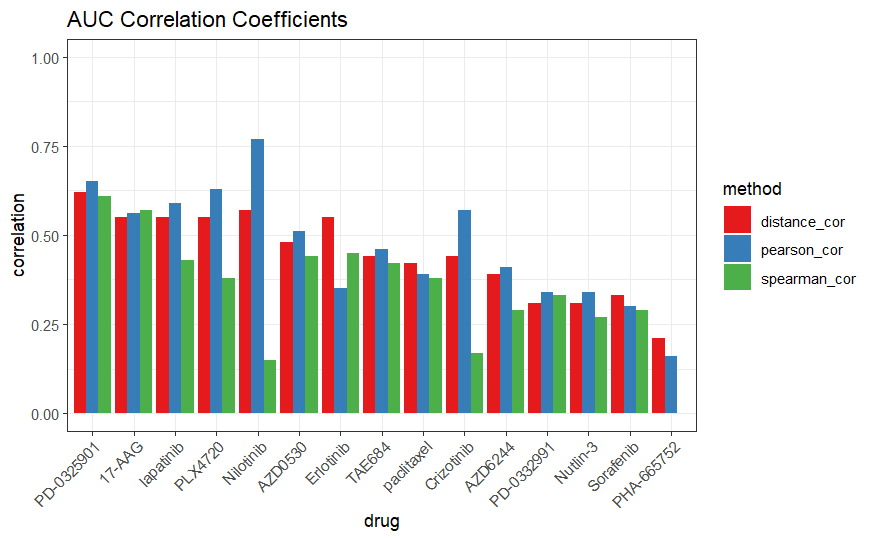
**theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +**

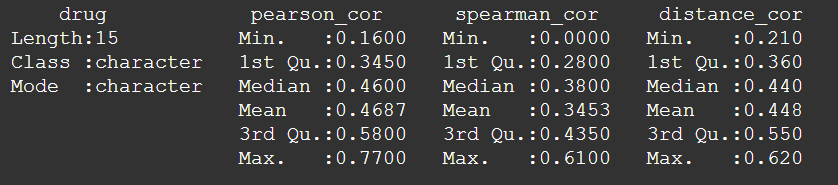
**ylim(0, 1) +**

**ggtitle("AUC Correlation Coefficients")**

**# Gorgeous :')**

**```**

****



ggplot(aes(x = drug, y = correlation, fill = method, group = method)) + geom\_bar(stat = "identity", position = position\_dodge()) + scale\_fill\_brewer(palette = "Set1") + theme(axis.text.x = element\_text(angle = 45, hjust = 1)) + ylim(0, 1) + ggtitle("AUC Correlation Coefficients")

First, we need to understand the purpose of each correlation measure. Pearson correlation focuses more on numerical data itself and how they differ in terms of a linear relationship, whereas Spearman correlation demonstrates the correlation between the trends of values. That is, if one variable increases and the other also increases, it would have a correlation of 1. For data that is not strictly linear, Spearman correlation is better because it takes into account increases and decreases in values, whereas Pearson correlation smooths out these changes.